EPFL-Smart-Kitchen: An Ego-Exo Multi-Modal Dataset for Challenging Action and Motion Understanding in Video-Language Models

Andy Bonnetto[1,*], Haozhe Qi[1,*], Franklin Leong[1], Matea Tashkovska[1], Mahdi Rad[2], Solaiman Shokur[1,3], Friedhelm Hummel[1,4,5,6], Silvestro Micera[1,3], Marc Pollefeys[2,7], Alexander Mathis[1,✉*]

1: École Polytechnique Fédérale de Lausanne (EPFL), Lausanne
2: Microsoft 3: Scuola Superiore Sant'Anna, Pisa
4: Swiss Federal Institute of Technology Valais (EPFL Valais), Sion
5: Clinique Romande de Réadaptation, Sion
6: University of Geneva Medical School, Geneva
7: Eidgenössische Technische Hochschule (ETH), Zürich

Abstract

Understanding behavior requires datasets that capture humans while carrying out complex tasks. The kitchen is an excellent environment for assessing human motor and cognitive function, as many complex actions are naturally exhibited in kitchens from chopping to cleaning. Here, we introduce the EPFL-Smart-Kitchen-30 dataset, collected in a noninvasive motion capture platform inside a kitchen environment. Nine static RGB-D cameras, inertial measurement units (IMUs) and one head-mounted HoloLens 2 headset were used to capture 3D hand, body, and eye movements. The EPFL-Smart-Kitchen-30 dataset is a multi-view action dataset with synchronized exocentric, egocentric, depth, IMUs, eye gaze, body and hand kinematics spanning 29.7 hours of 16 subjects cooking four different recipes. Action sequences were densely annotated with 33.78 action segments per minute. Leveraging this multi-modal dataset, we propose four benchmarks to advance behavior understanding and modeling through 1) a vision-language benchmark, 2) a semantic text-to-motion generation benchmark, 3) a multi-modal action recognition benchmark, 4) a pose-based action segmentation benchmark. We expect the EPFL-Smart-Kitchen-30 dataset to pave the way for better methods as well as insights to understand the nature of ecologically-valid human behavior. Code and data are available at https://amathislab.github.io/EPFL-Smart-Kitchen.

1 Introduction

Understanding human behavior is fundamental across multiple domains - from augmented reality[9] and robotics [18] to neuroscience [52, 54] and neuroengineering [56]. While we have made significant progress in behavioral analysis through action recognition [89, 37, 83, 17], action segmentation [86, 45, 98, 77] and motion generation [80, 99, 27], critical gaps remain. Current datasets face a fragmentation problem (Table 1). Existing datasets excel in isolated aspects of behavioral capture, but lack integration. Some datasets advance full-body 3D pose estimation but provide insufficient hand tracking for complex movements. Others offer detailed finger articulations but are limited to constrained environments, missing the crucial full-body context including the global

*A.B. and H.Q. contributed equally. Correspondence: alexander.mathis@epfl.ch

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

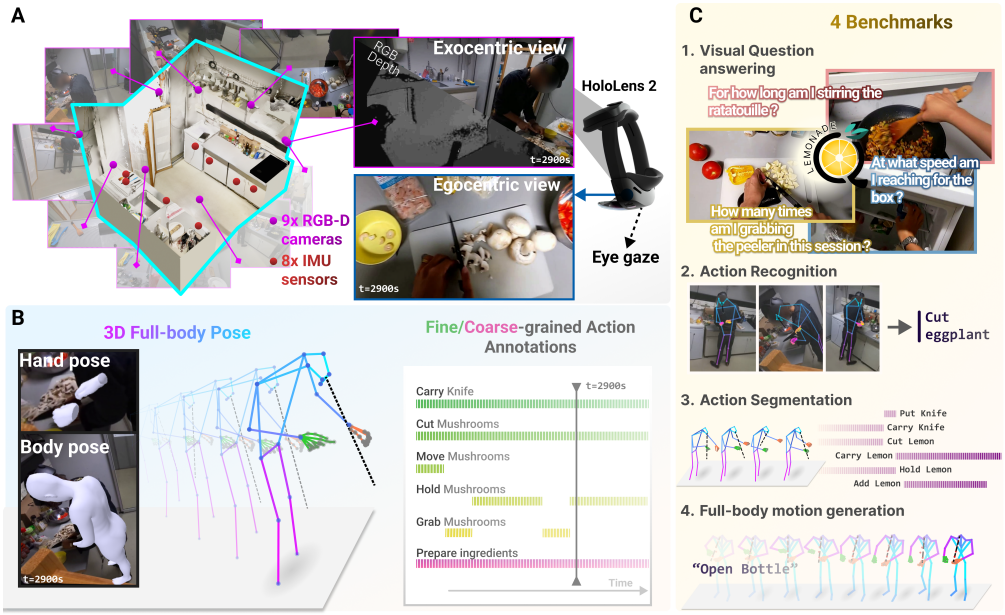

Figure 1: **The EPFL-Smart-Kitchen-30, dataset and benchmarks.** (A) **Collected data**. 3D kitchen reconstruction, purple points are fixed RGB-D cameras. Subjects cook with a HoloLens 2 headset recording both egocentric videos and eye gaze. (B) **Extracted data**. 3D body and hand poses are extracted from multiple data sources. Fine-grained and coarse-grained action segments are densely annotated. (C) **Benchmarks.** We propose four benchmarks based on the EPFL-Smart-Kitchen-30 dataset. A Visual question answering benchmark, an action segmentation benchmark, an action recognition benchmark and a full-body motion generation benchmark.

position. Moreover, some datasets fail to capture two essential components of natural behavior: goal-directed actions and eye movements. Human actions are inherently purposeful and guided by visual attention [28], yet most current datasets do not incorporate these elements, resulting in an incomplete representation of behavior. Comprehensive datasets of full-body, including hand and eye tracking alongside synchronized multi-view video and detailed, hierarchical action annotations are currently missing, and significantly hinder our ability to analyze natural human behavior [28, 52, 77].

We present the EPFL-Smart-Kitchen-30, a dataset that captures humans in authentic cooking scenarios with multimodality. It features both egocentric and exocentric perspectives through ten synchronized camera views, providing excellent visual coverage of natural cooking behaviors. EPFL-Smart-Kitchen-30 includes multiple modalities: RGB and depth images, IMU data, eye gazes, and 3D hand/body poses (Figure 1A-B).The EPFL-Smart-Kitchen-30 compares favorably to other datasets (Table 1) and promises to advance multimodal fine-grained action understanding. The scale is substantial: 29.7 hours of multi-view, multimodal recordings from 16 participants across 49 complete cooking sessions, from recipe reading to cleanup. The dataset defines 763 fine-grained actions, ensuring dense, hierarchical action annotation and exclusive action definitions. Sessions are densely annotated, yielding 55,361 fine-grained action segments and 4,828 coarse-grained activity segments—about 33 actions per minute.

With the annotated data, we build four benchmarks for action understanding and modeling (Figure 1C). First, we introduce Lemonade, a novel approach that transforms our ground truth annotations and pose estimations into challenging close-ended question-answer pairs (QA). This benchmark specifically tests the behavioral understanding capabilities of video-language models (VLMs). Second and third, we propose action recognition and segmentation benchmarks that span multiple modalities, providing empirical insights into procedural human behavior. Fourth, we present a full-body motion generation benchmark that highlights EPFL-Smart-Kitchen-30's unique value for generative tasks, demonstrating how our integrated data approach enables more natural and contextually appropriate motion synthesis. In summary, we make the following contributions (Figure 1):

- We capture 30 hours of goal-directed cooking behavior from ego-exo perspectives

- We densely annotate fine-grained actions and coarse-grained activities.
- We propose multimodal behavior understanding (action recognition, segmentation and vision-language question answering) and modeling benchmarks

These contributions collectively address the fragmentation problem in behavioral analysis and provide the research community with new possibilities for integrated, context-rich human behavior understanding.

2 Related work

2.1 Datasets of human behavior

Many datasets have been proposed that record participants executing purposeful motions [72, 47, 65], which were further extended to fitness activities by datasets like EgoExo-Fitness [41] and FLAG3D [78] to include more complex human body motions. Traditional motion capture approaches focused on isolated movements, offering high controllability but sacrificing critical contextual information. Recent research has shifted toward recording behavior in natural settings, enabling the study of authentic transitions and sequence patterns that characterize genuine human activity. In particular, absolute positioning determines the agent's spatial location and facilitates the analysis of its interactions with the environment [49]. Assembly-based datasets collect structured object interactions [6, 91, 3, 70], but by using a greater variety of actions and natural environments, cooking is getting popular for building action datasets such as EPIC-KITCHENS-100 [14], Humans in kitchens [79] or certain sequences of the large EgoExo4D [24] dataset. The EPFL-Smart-kitchen-30's dense annotations suit the characterization of body and hand movement transitions and distinguish themselves with their unambiguous and rich action descriptions (Table 1).

Language can flexibly describe behavior, and VLMs promise to capture that richness. The general video understanding of VLMs has been evaluated with exhaustive benchmarks such as MVBench [38] and Video-MME [20]. More specific challenges subsequently developed to tackle long-term understanding [105, 51] and egocentric video understanding [51, 30]. In the case of behavior understanding, ActivityNet-QA [96] and NExT-QA [94] evaluate the causal and temporal abilities of VLMs. While subsets of certain benchmarks [13, 38, 105] contain questions related to motion, they mostly focus on understanding behavior at the event level. EPFL-Smart-Kitchen-30 enables a fundamentally different approach. Our Lemonade benchmark introduces questions that specifically probe the understanding of human kinematics and fine-grained behavioral details that previous datasets simply cannot address.

Table 1: **Action dataset comparison.** # indicates "number of" for simplicity. The remaining columns mark following features: parametric model for motion representation (PM), structured actions (SA), markerless video recording (ML), depth recording (DR), and absolute positioning (AP). AP refers to global positioning (derived from point clouds) that is consistent across different sessions. Note: EgoExo4D [24] reports 1422h by summing per camera recording time, where the total activity is 180h, yet only 88.8h annotated with MSCOCO keypoints (numbers from [49]).

	Datasets	Total hours	Duration (min)	# segments	Seg. per min	# action classes	# Ego/ Exo	Body PM	Hand PM	Eye Gaze	SA	ML	DR	AP
Video-focused	Meccano [67]	6.9	20.7	8,857	21.4	61	0/1	✗	✗	✓	✓	✓	✓	✗
	IKEAASM [6]	11.7	1.9	17,577	8.4	33	0/3	✗	✗	✗	✓	✓	✓	✓
	EPIC-100 [14]	100.0	8.6	89,977	15.0	4,053	1/0	✗	✗	✗	✓	✓	✗	✗
	EgoExo4D [24]	180	15.3	20,406	4.5	689	1/4-5	✗	✗	✓	✓	✓	✗	✗
	HoloAssist [91]	166.0	4.5	184,838	18.6	1,887	1/0	✗	✗	✓	✓	✓	✓	✓
	EgoExo-Fitness [41]	32	1.5	6,131	4	12	0/4	✗	✗	✗	✓	✓	✗	✗
Motion-focused	AMASS [50]	43	0.22	11,451	0.22	-	-	✓	✓	✗	✗	✗	✗	✗
	BABEL [65]	43	0.39	28,000	10.7	250	-	✓	✗	✗	✓	✗	✗	✗
	HumanML3D [26]	28.6	0.12	14,616	8.5	-	-	✓	✗	✗	✗	✗	✗	✗
	HumanAct12 [25]	-	-	1,191	-	34	-	✓	✗	✗	✓	✓	✓	✗
Video-motion focused	Assembly101 [70]	41.8	7.1	84,460	33.1	1,380	4/8	✗	✗	✗	✓	✓	✗	✓
	H2O [34]	5.5	0.33	1,000	3.0	36	1/4	✗	✓	✗	✓	✓	✓	✗
	MotionX [43]	144	0.11	81,100	9.4	-	0/1	✓	✓	✗	✗	✓	✗	✗
	Nymeria [49]	300	15	-	-	-	1/1	✓	✗	✓	✗	✗	✗	✓
	EPFL-Smart-Kitchen-30	29.7	35.9	60,189	33.78	768	1/9	✓	✓	✓	✓	✓	✓	✓

By leveraging our multimodal data integration, we can evaluate models on their ability to reason about body movements and hand-object interactions (Figure 4).

2.2 Models for behavior understanding

The ability to predict movement patterns provides a valuable approach to understanding behavior. Movement can be captured in various forms, including video recordings, pose estimation data, and IMU recordings. Improvements in deep learning models together with their increase in computational power levels have led to the development of many multi-view, multimodal action understanding algorithms [71, 88, 104, 7, 93, 73]. Shah et al. [71] leverage contrastive learning to align the feature spaces from different views. Wang et al. [88] use an adversarial generative network to constrain RGB and depth modality information. HandFormer [73] combines 3D hand poses and RGB frames together for action recognition. LaViLa [104] learns video representations from pre-trained large language models. TIM [7] designs time interval encodings to incorporate visual and audio events. Despite progress, current methods are limited in views and modalities, partially due to the lack of large-scale multi-view, multimodal action datasets. With our EPFL-Smart-Kitchen-30 dataset, we set up multi-view, multimodal action understanding benchmarks taking and comparing exocentric videos, egocentric videos, full-body pose estimations, and eye gaze modalities as input, with the possibility to also include depth videos and IMU recordings.

Another approach for behavior understanding is through the ability to generate movement of a target behavior. Recently, text-to-motion generation gained a lot of attention [80, 8, 99, 27, 81, 63, 101]. We propose a novel semantic text-to-motion generation benchmark that considers full-body pose representations, including eye gaze. for situated motion generation. This contrasts with the commonly used KIT [64] and HumanML3D [26], which do not incorporate hand models or gaze information.

By integrating language, VLMs provide more flexible ways to understand behavior. VideoLLaMA3 [97] captures fine-grained details and temporal dynamics in videos through its dynamic resolution mechanisms and advanced positional embedding strategies, whereas Qwen2.5-VL [5] and Intern VL2.5 [11] better integrate multimodal inputs. Specific tasks such as long-term video understanding usually rely on video compression [36, 74, 40, 42] or on extending their context length [102, 92, 10, 46]. We challenge these models to operate beyond their conventional performance by proposing a benchmark that leverages behavioral context and kinematics.

3 The EPFL-Smart-Kitchen-30

Here we introduce the EPFL-Smart-Kitchen-30 dataset, which features multi-view, multimodal data of human cooking with fine-grained and coarse-grained action annotations (Figure 1B). We will

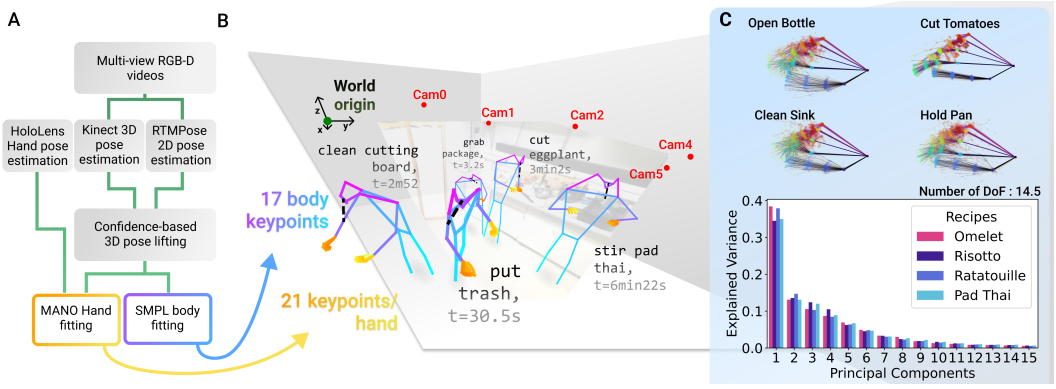

Figure 2: **Full-body 3D pose estimation** (A) Pipeline for 3D pose estimation (B) Poses and camera positions are defined relative to a global coordinate system comparable in the same environment. We illustrate several reprojected 3D poses on camera 6. (C) Characterization of right-hand poses to show the captured kinematic diversity: (Top) examples of hand poses for four actions, (Bottom) Number of principal components necessary to capture the right-hand poses during cooking for the Exo-Hand in each recipe, around 14 degrees of freedom (DoF) are necessary to explain 95% of the variance.

describe the setup (Sec. 3.1) and the data collection procedure (Sec. 3.2). Then, we illustrate the 3D pose regression (Sec. 3.3) and detail the action annotation characteristics (Sec. 3.4).

3.1 The EPFL-Smart-Kitchen setup

Capturing multi-view, multimodal data is a challenging task that requires the synchronization and calibration of multiple sensors. To capture naturalistic cooking behaviors, we built the EPFL-Smart-Kitchen, a fully functional kitchen with appliances and utensils. Cooking materials, including pots, pans, and other utensils, were provided to the subjects along with the ingredients and spices necessary for preparing the recipe (Supp. Sec. A).

To minimally affect the subjects' natural movements while capturing multimodal information, we installed nine Microsoft Kinect Azure RGB-D cameras [57] at strategic points inside the kitchen, four focusing on the global exocentric view and five focusing on local exocentric views (counters, stove, and sink, see Figure 1A). We additionally equipped the kitchen with eight IMU sensors on the frequently used equipment (e.g., fridge door, five cupboard doors, knife, and spatula). Subjects wore a Hololens 2 headset [85], a mixed-reality headset that can capture egocentric views and eye gaze data under global calibration. We synchronized all devices using audio signals and a trigger, and calibrated all the cameras (Supp. Sec. A.2).

3.2 Data collection procedure

To capture realistic cooking scenarios, subjects prepared a meal from reading a recipe to cleaning up, leading to significantly longer recordings than in most existing action datasets (Table 1). We recruited 16 subjects (four males and twelve females, two left-handed, ages 20-46) to cook for up to five sessions in the EPFL-Smart-Kitchen (Supp. Sec.A.3). During each session, subjects are asked to follow one of four different recipes (omelet, pad thai, risotto, ratatouille), adapted to their preferences and requirements. Overall, we recorded and processed 29.7h of cooking experiments, corresponding to 3,207,600 frames per camera for 49 cooking sessions. All procedures were approved by the EPFL-Ethical Board. Subjects' faces are anonymized across all videos to address privacy concerns. All subjects consented and were informed about the ethics (Supp. Sec. D.1).

3.3 Estimation of 3D motions

We placed the cameras so that both the body and the hands of the participants are visible from at least three angles. Four cameras captured global body information, while five cameras captured local hand information. Using multi-view RGB-D video, we conduct body/hand mesh fitting and tracking using all 10 camera views, extracting 2D and 3D pose information from each view with existing pose estimation tools. Specifically, we extract 2D body and hand poses using RTMPose [31], available in DeepLabCut v3 [53], 3D body poses and tracklets using the Kinect body tracking SDK [58], and 3D hand poses using the HoloLens 2 hand tracking toolkit [59]. We lift 2D poses to 3D poses and fit the SMPL [62] body mesh by minimizing the 3D joint, 2D reprojection, temporal smoothing, and regularization loss, as well as the hand 3D joint loss to fit the MANO hand mesh [68] (Figure 2A and Suppl. Sec. C). The average absolute error compared to triangulated manually-annotated 2D poses is $6.22cm \pm 5.16cm$ and $3.30cm \pm 5.12cm$ for the body and hand respectively, our margins are comparable with those of [24, 34] (Supp. Sec. C.5).

To illustrate the richness of the captured movement data (Figure 2B-C), we estimate the number of degrees of freedom (kinematic synergies) in pose space based on a common method in neuroscience [82, 69, 12]. We found that the dataset exhibits a large number of degrees of freedom (Figure 2C), which foreshadows the potential for studies on human behavior.

3.4 Annotation of fine-grained actions and coarse activities

The annotated action classes were defined with the following considerations. Firstly, many contemporary datasets (e.g.,[14, 91]) tend to allow the annotators to freely describe the actions and then post-hoc group the actions based on action similarity. This might lead to different names for similar actions (e.g., *pour* and *fill* [14]) and thus introduce ambiguity. We instead curated a set of verbs and nouns. We annotated with temporal overlaps between actions. For example, when labeling *cut tomato*, we also label *carry knife* and may label *hold tomato*. This enriches the annotation at a given

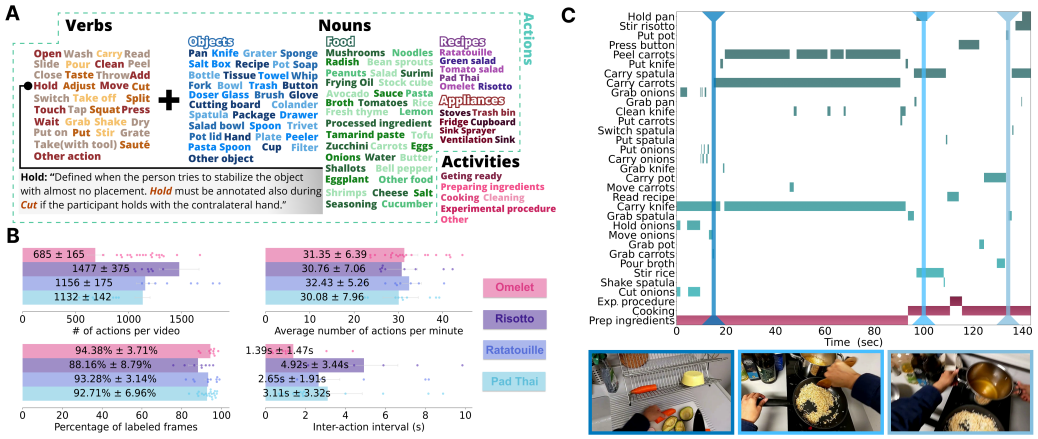

Figure 3: **Hierarchical action annotations**: (A) List of verbs, nouns and activities used for action annotation, each action (verbs) is specifically defined as shown for "Hold". (B) Statistics of annotation segments for fine-grained actions. (C) Example ethogram with selected egocentric frame to illustrate the richness of the actions (turquoise) and activity (pink) annotations comprising short and long segments that can overlap.

timestep and attempts to reduce ambiguity. Based on the above rationale, we define 33 verbs and 79 nouns, which compose 763 fine-grained actions. Each verb is defined by a rule-based description intended to prevent confusion (see example in Figure 3A and Supp. Sec. B). During the annotation procedure, we asked annotators to watch videos and annotate the start and stop times of actions. To define the behavioral contexts for each action, six coarse-grained exclusive activities were annotated, summing up to 4,828 segments. Thus, behavior is annotated in a hierarchical fashion [4, 77, 21].

The quality and reliability of annotations were validated following the protocol outlined in Supp. Sec. B.5 In total, 60,189 action segments were annotated, resulting in 33.78 action segments per minute (Figure 3B). The richness of the action annotation is demonstrated by a large variety of action lengths (from 1 second to 100 seconds, Figure 3C). Overall, the different views and modalities contribute to fine-grained action understanding in different aspects: 1) RGB frames focus on coarse-grained information while depth frames rather focus on geometric aspects; 2) the egocentric view and the eye gaze data captured from the HoloLens 2 are related to (part of) what the subject sees; 3) global exocentric views and the body poses capture the overall context; 4) local exocentric views, hand poses, and tool IMU data capture fine-grained movements and hand-object interactions.

4 Multimodal action and motion understanding benchmarks

Cooking involves many different actions that are sequenced in a goal-directed fashion to achieve a tasty outcome. In each experimental session, subjects go from reading the recipe and preparing the ingredients to creating the dish and ultimately cleaning up. To make progress towards analyzing such complex human behavior, we created four behavior understanding benchmarks (Figure 1C). Two benchmarks focus on multimodal behavior analysis: action recognition and action segmentation. These benchmarks are complemented by a full-body behavior synthesis benchmark (motion generation). Furthermore, we designed a question-answering benchmark (Lemonade) to understand human cooking behavior. Lemonade is structured for zero-shot evaluation. For the other benchmarks, we split the sessions into train, validation, and test sets. The training/validation sets are split into 26/7 sessions chosen so that every recipe is present in the validation set and to balance the number of rare action segments in both sets. The test set is composed of 16 sessions which also include new subjects. The curated dataset used for the benchmark excludes actions with less than 3 instances and is composed of 31 verbs, 78 nouns, and 581 actions together with six activities.

4.1 Lemonade: Language models Evaluation of MOtion aNd Action-Driven Enquiries

Rationale. VLMs exhibit remarkable potential for understanding human behavior [90, 100, 36, 95]. They raise intriguing questions: Can they accurately predict preceding or subsequent actions in behavioral sequences? Are they able to infer long-term behavioral patterns from just a few frames?

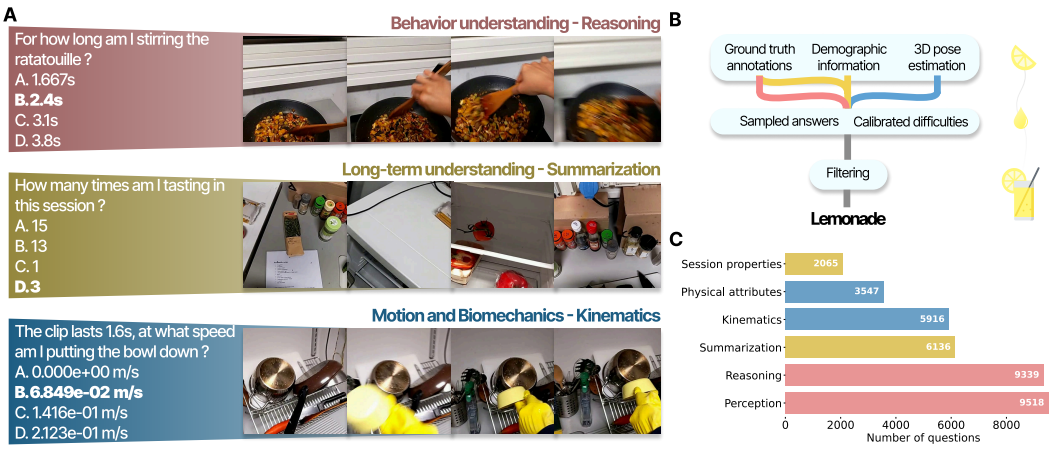

Figure 4: **Lemonade**: (A) Examples of video question pairs for each category. More examples in Supp. Sec. F.1 (B) Questions are designed from ground truth annotations. (C) Distribution of questions for all subcategories.

Furthermore, can their general knowledge enable precise distance and velocity estimations from video data alone? EPFL-Smart-Kitchen-30 provides the ideal testbed to explore these fundamental questions about machine understanding of natural human behavior (Figure 4A). Thus, we introduce Lemonade: **L**anguage models **E**valuation of **MO**tion a**N**d **A**ction-**D**riven **E**nquiries. The Lemonade framework (Figure 4B) is designed to generate millions of unique QA pairs by combining various video clips, question formats, and answer types. Lemonade consists of 36,521 closed-ended QA pairs linked to egocentric video clips, categorized into three groups and six subcategories (Figure 4C). 18,857 QAs focus on behavior understanding, leveraging the rich ground truth behavior annotations of the EPFL-Smart-Kitchen to interrogate models about perceived actions (Perception) and reason about unseen behaviors (Reasoning). 8,201 QAs involve longer video clips, challenging models in summarization (Summarization) and session-level inference (Session properties). The remaining 9,463 QAs leverage the 3D pose estimation data to infer hand shapes, joint angles (Physical attributes), or trajectory velocities (Kinematics) from visual information. More examples and details on QA design can be found in the Supp. Sec. F.1

Baselines and Metrics. Based on state-of-the-art results from other recent benchmarks [84], we evaluated three open-source and one closed-source VLM SoTA models, namely InternVL2.5 [11], LLaVA-OneVision [36], Qwen2.5-VL [5] and Gemini 2.0 Flash [15]. To ensure consistent evaluation and enable future comparisons, all models are evaluated using lmms-eval [100], where Lemonade is implemented as a new evaluation task. To interpret the results, we manually answered 1,662 questions. As is commonly done for question answering, we evaluate the model performance as average accuracy [20, 38, 94, 96].

Results. Different VLMs achieved high accuracy in identifying ongoing actions and activities, as well as predicting immediate next and previous actions. However, Lemonade exposed critical limitations in VLMs in predicting general context information, distances, timings, and body kinematics (Table 2). However, this benchmark is also challenging for humans. Merely relying on visual input and language proved insufficient for these precise kinematic estimations, which require accurate frame timing and depth reference data. To overcome these challenges, future models could benefit from explicitly integrating additional modalities, which is becoming possible with multimodal language models [48, 19].

4.2 Action recognition benchmark

Rationale. Given a trimmed action segment, the action recognition model needs to predict the corresponding action class [89, 37, 83, 17]. We built a fine-grained action recognition benchmark for 763 classes with a long-tailed distribution and allow different data types as input. Specifically, we formulated flexible masked auto-encoding baselines taking the egocentric view, one exocentric view, the 3D body poses, the 3D hand poses, and the eye gaze rays as input and compared different combinations of those data sources.

Table 2: Accuracy of VLMs (8 frames) on the lemonade benchmark per category; chance level is 25%. Human carried out answered 1,662 samples. Detailed results in Supp. Table. F.2

Models	Perc.	Reas.	Sess.	Summ.	Phys.	Kin.	**All**
Human*	62.32	52.38	62.38	44.40	70.78	38.34	54.75
Gemini-2.0-Flash	41.47	41.03	**49.10**	35.09	**40.26**	21.69	37.39
LLaVa-OneVision	**47.23**	41.46	36.76	41.87	34.03	**34.09**	**40.85**
InternVL2 5-8B	43.94	41.52	45.76	33.80	39.89	27.74	38.70
Qwen2.5-VL-32B (4fr.)	43.72	**45.87**	46.25	**42.75**	32.22	28.52	40.67
Qwen2.5-VL-7B	42.86	43.58	48.14	37.66	38.93	25.64	39.30

Baselines and Metrics. We adapted VideoMAE [83, 21] model into a multi-modal MAE, enabling it to take multi-view videos, 3D poses, and eye gazes as inputs (Supp. Sec. F.4). Like others, we evaluate the model performance by Top-1 and Top-5 accuracy for verb, object, and action classes [91, 70]. Considering that the number of action samples across different actions has a long-tail distribution, we selected the top 180 most frequent actions as head actions and report the performance for both head actions and tail actions separately.

Results. Simple concatenation of all modalities as inputs yielded slightly better performance compared to single video modality (Table 3). Meanwhile, transfer learning from a ViT model trained on EPIC-KITCHENS-100 [14] boosted the performance for baselines with visual inputs and reduced the performance for pose only. Furthermore, adding (hand and body) pose information to the video-based models yielded better overall performance, which mainly benefits from the verb prediction improvements. However, simply concatenating tokens from different modalities barely improved the performance. To efficiently utilize multiple modalities without significant computational cost, we integrate egocentric view, multi-exocentric views, and hand pose data (⚙) (Supp. Sec. F.4.4) to boost the performance by 21.6% when trained from scratch (Supp. Table F.3) and 6.3% when trained from the pretrained model, over the egocentric-only model. Overall, we hope this benchmark will inspire the community to create models that can effectively use multiple modalities.

Table 3: **Fine-grained action recognition benchmark results from pretrained model.** 📷: egocentric view, ▶: global exocentric view, 🧍: 3D body pose, ✋: 3D hand pose, 👁: eye gaze, ⚙(✋×multiple ▶) : hand cropped videos. Combining modalities has the potential to increase the performance. Our best results are achieved by cleverly merging these modalities together.

Modalities	All Classes Accuracy Top1/5			Head Classes Accuracy Top1/5			Tail Classes Accuracy Top1/5		
	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun
📷	37.51/62.94	57.72/92.18	52.03/79.05	41.12/67.00	59.74/93.36	55.62/82.11	16.64/39.51	46.06/85.35	31.27/61.38
📷🧍	37.87/63.56	58.90/93.59	52.56/79.64	41.55/67.67	60.56/94.66	56.43/82.84	16.66/39.85	49.28/87.44	30.19/61.11
📷▶	37.57/63.13	58.38/92.92	52.29/78.45	41.14/67.31	60.43/93.90	55.90/81.62	16.97/39.00	46.54/87.24	31.46/60.15
🧍✋	11.80/25.49	38.67/78.80	19.83/42.23	13.55/28.77	39.87/80.65	22.31/46.29	1.70/6.57	31.77/68.15	5.49/18.79
📷🧍✋	38.31/64.75	60.41/93.78	52.51/79.91	41.83/68.89	61.96/94.79	56.27/83.14	17.97/40.90	51.45/87.94	30.81/61.27
📷🧍✋👁	37.35/62.34	60.78/93.04	50.58/77.30	40.91/66.05	62.20/93.92	54.15/80.48	16.85/40.97	**52.55**/87.94	30.02/58.99
📷▶🧍✋👁	37.49/62.76	**61.04**/93.59	50.94/77.52	41.09/66.50	62.53/94.51	54.75/80.57	16.66/41.21	52.42/**88.29**	28.95/59.91
📷⚙(✋× ▶)	**40.03/67.01**	60.80/**94.65**	**55.38/82.58**	**43.60/71.25**	**62.60/95.76**	**59.00/85.62**	**19.44/42.52**	50.41/88.25	**34.48/65.02**

4.3 Action segmentation benchmark

Rationale. Given an untrimmed video, action segmentation requires the model to predict one or multiple action classes for every frame [86, 45, 98, 77]. Given the absence of popular (and comprehensive) action segmentation benchmarks from 3D pose data (Section 2), we built an action segmentation benchmark that compares the impact of different input data (body, hand, eyes, video features). One might expect that actions such as moving through the kitchen will be better predicted from the body, while motions like cutting require hand pose keypoints. Therefore, we used combinations of body pose (🧍), hand poses (✋) and eye gazes (👁) to form the input as they are computationally more efficient than deep visual features. We additionally compared the performance when using video features from VideoMAE as input.

Baselines and Metrics. We consider state-of-art pose estimation models proposed in DLC2Action [32] to perform action segmentation. The toolbox adapted state-of-the-art models for

RGB-based action segmentation tasks (Breakfast [33] and 50Salads [76]), such as MS-TCN++ [39], EDTCN [35], and C2F-TCN [75], to work directly on pose estimation data (vs deep visual features): MS-TCN3 and C2F-Transformer. We used kinematic features and VideoMAE [83] features as input to the models (Supp. Sec. F.6.3). Each action is evaluated separately using standard metrics in action segmentation (Frame-wise F1, F1@50, edit distance) and ultimately averaged over action groups. All models were trained and evaluated using the DLC2Action toolbox [32].

Results. We observed that the benchmark is challenging for current action segmentation algorithms (Table 4 and Supp. Table F.4). Exo-Body performs similarly to Exo-Hand with a slight improvement for Exo-Hand albeit for different behaviors. We note that video information provided a boost in performance. These baselines showed an F1-score of 35.2% for verbs and 35.0% for nouns, highlighting significant potential for future advancements.

Table 4: **F1 scores for action segmentation benchmark.** 🧍: 3D body pose, ✋: 3D hand pose, 👁: eye gaze, 📷: egocentric view. *models modified to use pose as input data instead of image features.

	Verbs						Nouns						Activity					
	🧍	✋	👁	🧍✋	🧍👁	🧍✋📷	🧍	✋	👁	🧍✋	🧍👁	🧍✋📷	🧍	✋	👁	🧍✋	🧍👁	🧍✋📷
MS-TCN3	18.1	20.2	11.7	20.9	21.1	30.1	10.6	13.4	7.6	15.6	11.3	31.2	51.8	58.6	31.9	54.4	58.5	**72.9**
C2F-TCN*	18.8	20.1	12.2	22.1	22.2	34.6	12.0	14.3	7.9	16.1	10.8	**35.2**	54.5	55.4	41.3	61.8	61.2	72.2
C2F-Transf.	19.9	22.4	13.1	22.8	22.2	**35.0**	11.1	12.9	7.8	13.4	9.2	29.0	51.2	56.9	38.8	62.1	59.9	70.5
EDTCN*	19.6	23.0	11.9	22.1	25.2	34.3	11.9	11.2	7.1	12.3	11.9	24.3	49.0	53.5	32.0	53.1	54.2	71.0

4.4 Situated full-body motion generation benchmark

Rationale. With the diverse actions and motions in EPFL-Smart-Kitchen-30, our motion generation benchmark has three key innovations over the commonly used KIT [64] and HumanML3D [26] benchmarks. Existing motion generation benchmarks mainly focus on broad daily activities, sports, and dance, whereas the EPFL-Smart-Kitchen-30 contains hundreds of fine-grained behaviors. We extend motion generation beyond the body to include hands and eye gaze, defining it as full-body motion generation. Additionally, we provide egocentric visual features alongside action text as the condition, allowing situated motion generation.

Pre-processing, Baselines and Metrics To achieve robust full-body motion representation, we combine joint locations and angles for the body, hands, and eye gaze, resulting in a 327-dimensional redundant motion representation. We process fine-grained action text with linguistic tags and extract egocentric visual features with CLIP's text and image encoders [66]. As baselines, we adapt three strong motion generation models, T2M-GPT [99], MARDM-SiT [55] and MoMask [27], training them with verb-noun pairs and verb-only prompts (Supp. Sec. F.7). Like in HumanML3D [26], we train the quantitative evaluator on EPFL-Smart-Kitchen-30 to measure the R-Precision top-1 to top-3 (T1 to T3), multimodal distance (MMd), Fréchet Inception Distance (FID), Diversity (DIV), and Multimodality (MM).

Results. Our qualitative analysis (from MoMask [27]) revealed compelling examples of successfully generated full body motion sequences that accurately represent natural cooking behaviors (Figure 5). Quantitatively, we found that just like on HumanML3D[26], MoMask [27] consistently outperforms T2M-GPT [99] and MARDM-SiT [55] in different settings and metrics (Table 5), possibly due to the hierarchy and quantized tokenization way working better for the high-dimensional properties of the

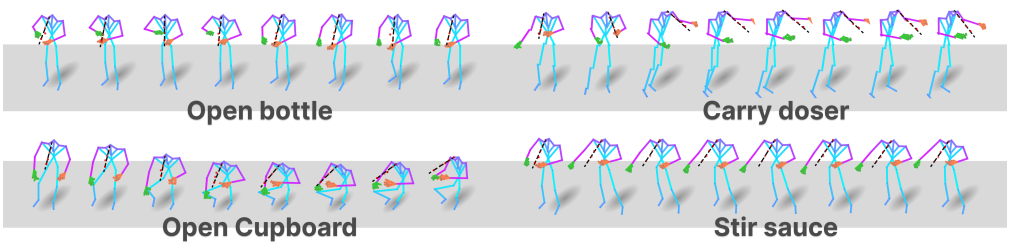

Figure 5: **Qualitative motion generation samples** from MoMask trained with Verb-Noun action pairs. The model creates realistic hand, body and eye gaze (dashed line).

full-body representation. Furthermore, when conditioning on the egocentric view, the model was able to find some clues to minimize the global distribution, and thus make the FID score lower.

Table 5: **Full-body motion generation results**. Models trained and evaluated on the EPFL-Smart-Kitchen-30. FID:Fréchet Inception Distance, DIV:Diversity, MM:Multimodality, MMd:Multimodal distance.

Condition	Vocab.	Model	T1 ↑	T2 ↑	T3 ↑	FID ↓	DIV ↑	MM ↑	MMd ↓
Text	Verb	T2M-GPT	0.254	0.432	0.567	2.640	7.288	1.555	3.379
		MARDM-SiT	0.262	0.473	0.637	2.242	7.590	1.849	4.155
		Momask	0.306	0.508	0.652	3.124	**8.347**	**2.940**	**2.048**
	Actions	T2M-GPT	0.271	0.434	0.542	2.378	7.031	0.852	4.015
		MARDM-SiT	0.320	0.548	0.650	1.230	7.704	1.427	4.103
		Momask	**0.372**	**0.566**	**0.683**	0.930	7.859	1.641	3.407
Text-Image	Verb	T2M-GPT	0.174	0.295	0.387	2.415	6.558	0.822	4.524
		MARDM-SiT	0.187	0.355	0.466	1.498	6.579	1.581	4.691
		Momask	0.197	0.333	0.436	0.858	6.696	1.665	4.121
	Actions	T2M-GPT	0.243	0.398	0.506	1.982	6.633	0.717	4.125
		MARDM-SiT	0.255	0.469	0.597	0.917	7.176	1.322	4.725
		Momask	0.276	0.441	0.552	**0.627**	7.141	1.554	3.937

5 Conclusion, future work and impact

We collected 30 hours of RGB-D video from ten synchronized views, 3D pose data, and hierarchical action annotations in a calibrated kitchen environment. All participant data has been anonymized to protect privacy (Table 6). The dataset's multimodal nature and fine-grained annotations enable analysis of complex behavioral patterns, object interactions, and visual attention mechanisms during goal-directed activities (Figure 2C). Our work complements recent large-scale datasets such as EPIC-KITCHENS-100, EgoExo4D, and Humans in Kitchens [14, 79, 24] by providing integrated multimodal data streams within a controlled environment. While these datasets excel in scale and environmental diversity, EPFL-Smart-Kitchen-30 offers synchronized multi-view video, pose estimation, and eye tracking data that enables new research directions in multimodal behavior understanding.

The four benchmarks we propose—action recognition, action segmentation, motion generation, and video question answering—demonstrate both the potential and current limitations of existing models on fine-grained behavioral tasks. Future work could leverage additional modalities in our dataset (IMU data, depth information) and develop models that more effectively integrate multiple data streams for robust behavior understanding. Additionally, we are collecting data from older and non-healthy participants (stroke and amputee patients) for future release, aiming to improve treatments for subjects with neurological disorders [56]. This will also increase the demographic representation of our dataset. Overall, we share multi-view, multimodal action understanding, modeling, and video question answering benchmarks to leverage the potential of multi-modality for improving action understanding and to fuel foundation models. This is particularly interesting for emerging multi-modal models [48, 19, 60, 95].

Acknowledgments: We thank members of the Mathis Group for Computational Neuroscience & AI (EPFL) for their feedback throughout the project. This work was funded by EPFL, Swiss SNF grant (320030-227871), Microsoft Swiss Joint Research Center, and a Boehringer Ingelheim Fonds PhD stipend (H.Q.). We are grateful to the Brain Mind Institute for providing funds for hardware and to the Neuro-X Institute for providing funds for services.

Dataset and Code Release: Please check `https://amathislab.github.io/EPFL-Smart-Kitchen` for the latest updates.

Table 6: **Data and Code Availability**

Resource	**Location**
Code Repository	`https://github.com/amathislab/EPFL-Smart-Kitchen`
Dataset (Collected Data)	`https://zenodo.org/records/15535461`
Annotations (Pose & Behavior)	`https://zenodo.org/records/15551913`
Benchmark dataset and checkpoints	`https://huggingface.co/collections/amathislab/esk-benchmarks`

References

[1] Easymocap - make human motion capture easier. Github, 2021.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[3] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4575–4583, 2016.

[4] David J Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[6] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021.

[7] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. Tim: A time interval machine for audio-visual action recognition. *arXiv preprint arXiv:2404.05559*, 2024.

[8] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023.

[9] Yunqiang Chen, Qing Wang, Hong Chen, Xiaoyu Song, Hui Tang, and Mengxiao Tian. An overview of augmented reality technology. In *Journal of Physics: Conference Series*, page 022082. IOP Publishing, 2019.

[10] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024.

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[12] Alberto Silvio Chiappa, Pablo Tano, Nisheet Patel, Abigaïl Ingster, Alexandre Pouget, and Alexander Mathis. Acquiring musculoskeletal skills with curriculum-based reinforcement learning. *Neuron*, 112 (23):3969–3983, 2024.

[13] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024.

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[15] Google DeepMind. Introducing Gemini 2.0: Our new AI model for the agentic era, 2024.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2969–2978, 2022.

[18] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2): 230–244, 2022.

[19] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2093–2103, 2024.

[20] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024.

[21] Valentin Gabeff, Haozhe Qi, Brendan Flaherty, Gencer Sumbül, Alexander Mathis, and Devis Tuia. Mammalps: A multi-view video behavior monitoring dataset of wild mammals in the swiss alps. *arXiv preprint arXiv:2503.18223*, 2025.

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[23] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022.

[24] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023.

[25] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[26] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.

[27] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024.

[28] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4): 188–194, 2005.

[29] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint arXiv:2501.02955*, 2025.

[30] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *Advances in Neural Information Processing Systems*, pages 3343–3360. Curran Associates, Inc., 2022.

[31] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

[32] Elizaveta Kozlova, Andy Bonnetto, and Alexander Mathis. Dlc2action: A deep learning-based toolbox for automated behavior segmentation. *bioRxiv*, 2025.

[33] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[34] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021.

[35] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[37] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022.

[38] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024.

[39] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[40] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024.

[41] Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *European Conference on Computer Vision*, pages 363–382. Springer, 2024.

[42] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26679–26689, 2023.

[43] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[45] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10139–10149, 2023.

[46] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2025.

[47] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020.

[48] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, pages 417–435. Springer, 2022.

[49] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pages 445–465. Springer, 2025.

[50] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019.

[51] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, pages 46212–46244. Curran Associates, Inc., 2023.

[52] Antonella Maselli, Jeremy Gordon, Mattia Eluchans, Gian Luca Lancia, Thomas Thiery, Riccardo Moretti, Paul Cisek, and Giovanni Pezzulo. Beyond simple laboratory studies: Developing sophisticated models to study rich behavior. *Physics of Life Reviews*, 46:220–244, 2023.

[53] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[54] Mackenzie Weygandt Mathis, Adriana Perez Rotondo, Edward F Chang, Andreas S Tolias, and Alexander Mathis. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21):5814–5832, 2024.

[55] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*, 2024.

[56] Silvestro Micera, Matteo Caleo, Carmelo Chisari, Friedhelm C Hummel, and Alessandra Pedrocchi. Advanced neurotechnologies for the restoration of motor function. *Neuron*, 105(4):604–620, 2020.

[57] Microsoft. Azure kinect dk documentation. 2019.

[58] Microsoft. Azure kinect body tracking sdf. 2019.

[59] Microsoft. Hand tracking: Hololens mixed reality toolkit. 2022.

[60] David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In *Advances in Neural Information Processing Systems*, 2023.

[61] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[62] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[63] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024.

[64] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4 (4):236–252, 2016.

[65] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021.

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[67] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.

[68] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[69] Marco Santello, Martha Flanders, and John F Soechting. Postural hand synergies for tool use. *Journal of neuroscience*, 18(23):10105–10115, 1998.

[70] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[71] Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M de Melo, and Rama Chellappa. Multi-view action recognition using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3381–3391, 2023.

[72] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[73] Md Salman Shamil, Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. On the utility of 3d hand poses for action recognition. *arXiv preprint arXiv:2403.09805*, 2024.

[74] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

[75] Dipika Singhania, Rahul Rahaman, and Angela Yao. Iterative contrast-classify for semi-supervised temporal action segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2262–2270, 2022.

[76] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.

[77] Lucas Stoffl, Andy Bonnetto, Stéphane d'Ascoli, and Alexander Mathis. Elucidating the hierarchical nature of behavior with masked autoencoders. In *European Conference on Computer Vision*, pages 106–125. Springer, 2025.

[78] Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22106–22117, 2023.

[79] Julian Tanke, Oh-Hun Kwon, Felix B Mueller, Andreas Doering, and Juergen Gall. Humans in kitchens: a dataset for multi-person human motion forecasting with scene context. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc.

[80] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

[81] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[82] Emanuel Todorov and Zoubin Ghahramani. Analysis of the synergies underlying complex hand manipulation. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4637–4640. IEEE, 2004.

[83] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093, 2022.

[84] Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding, 2025.

[85] Dorin Ungureanu, Federica Bogo, Silvano Galliani, Pooja Sama, Xin Duan, Casey Meekhof, Jan Stühmer, Thomas J Cashman, Bugra Tekin, Johannes L Schönberger, et al. Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*, 2020.

[86] Beatrice van Amsterdam, Abdolrahim Kadkhodamohammadi, Imanol Luengo, and Danail Stoyanov. Aspnet: Action segmentation with shared-private representation of multiple data sources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2384–2393, 2023.

[87] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[88] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6212–6221, 2019.

[89] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[90] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[91] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023.

[92] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024.

[93] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[94] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021.

[95] Shaokai Ye, Haozhe Qi, Alexander Mathis, and Mackenzie W Mathis. Llavaction: evaluating and training multi-modal large language models for action recognition. *arXiv preprint arXiv:2503.18712*, 2025.

[96] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019.

[97] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[98] Junbin Zhang, Pei-Hsuan Tsai, and Meng-Hsun Tsai. Semantic2graph: Graph-based multi-modal feature fusion for action segmentation in videos. *arXiv preprint arXiv:2209.05653*, 2022.

[99] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023.

[100] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

[101] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023.

[102] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision, 2024.

[103] Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv preprint arXiv:2309.16669*, 2023.

[104] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023.

[105] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe the dataset, benchmark, and baselines.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Certain limitations are described in the last section (Sec. 5).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the data and the code to reproduce the results will be released upon acceptance. Our team has a history of providing widely used datasets and code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset is available on Zenodo at `https://zenodo.org/records/15535461`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We mention training and test details in the main text and also provide more details in the supplemental material

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We followed those guidelines for the statistics of our novel dataset, including the average duration of the sessions, pose variance analyses and action annotation statistics. Note that our benchmark performances do not rely on statistical tests due to the running cost (as it is common in ML).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We detail the used computational resources in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conducted research following the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We addressed this in the final section (Sec. 5). In brief, we do not think that there are negative implications, but many potentially positive ones as our work enables characterizing basic human cooking behavior, which is useful for assessing and improving motor function.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We will include a permissive license for academic research (likely CC BY-NC) along with our dataset upon acceptance. Following our institution's ethics guidelines, participants have been anonymized and gave consent to be remunerated participants.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models and datasets that we used are cited and credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper discusses the data collection process as well as the content of the dataset. The supplementary materials contain more information. Participants had to give consent, and signed a form explaining the purpose of the work. This procedure followed the IRB guidelines and was approved.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: All methods were approved by our Ethics board, subjects were remunerated, and consented. Details are contained in the supplementary text.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We did receive the mandatory IRB approval from our institution and have followed all guidelines.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The usage of LLMs in this paper is exclusive to the utilization of Video Language models for evaluation on the Lemonade benchmark.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.