

SUPPLEMENTARY MATERIALS OF HUMANPCR

Anonymous authors

Paper under double-blind review

1 DETAILS OF DATA SOURCES

1.1 SOURCE DATASET FULL LIST

All source datasets used in HumanPCR are listed in Table 1, including their licenses, websites, and descriptions.

Table 1: Details of the data sources employed.

Name	License	URL	Description
VSR Liu et al. (2023)	Apache License 2.0	https://github.com/cambridgelt1/visual-spatial-reasoning	Image-caption corpus for assessing whether spatial relationship statements accurately describe scenes
RefCOCOg Kazemzadeh et al. (2014)	Apache License 2.0	https://github.com/lichengunc/refer (for original RefCOCO)	Dataset linking long free-form referring expressions with their target objects or persons in COCO images
Motion-X Lin et al. (2023)	CC BY-NC-SA 4.0	https://motion-x-dataset.github.io/	Collection of expressive 3D whole-body motion sequences with SMPL-X annotations across diverse actions
HAKE-PaSta Li et al. (2022)	MIT License	http://hake-mvig.cn/	Provides part-state labels that capture fine-grained body and hand-object interactions
MEBOW Wu et al. (2020)	Only for research purposes	https://chenyanwu.github.io/MEBOW/	COCO-MEBOW supplies images annotated for single-view human body orientation estimation in the wild
GazeFollow Reichert et al. (2015)	Publicly available	http://gazefollow.csail.mit.edu/	Images labeled with gaze target regions so models can predict where each depicted person is looking
Clothing Attribute Dataset Chen et al. (2012)	Publicly available	https://purl.stanford.edu/tb980qz1002	Street-fashion image set annotated with clothing categories and visual attributes such as neckline or pattern
ChaLearn LAP Escalera et al. (2017)	The image copyright belongs to the original authors	https://chalearnlap.cvc.uab.cat/	Datasets from the ‘Looking at People’ challenges
MPII Human Pose An-driluka et al. (2014)	Simplified BSD License (Research purposes, non-commercial for images)	http://human-pose.mpi-inf.mpg.de/	Benchmark of everyday images with 16-joint annotations for articulated human pose estimation
CelebA Liu et al. (2015)	Non-commercial research use	https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	Face dataset with celebrity images annotated for binary attributes and landmarks, used for facial-attribute research

Continued on next page

Continuation of Table 1 (continued from previous page)

Name	License	URL	Overview
FlickrCI3D Fieraru et al. (2020)	CC-BY (images, ages)	(im- https://ci3d.imar.ro/)	Flickr images annotated with 3D body-region contact signatures for modeling social interactions
Hands23 Cheng et al. (2023)	CC-BY (ages)	(im- https://github.com/ddshan/hands23_data)	Egocentric images labeled with hand regions, contact areas and grasp types during everyday object manipulation
DECO pathi et al. (2023)	Tri-CC-BY (images, only from HICO and COCO)	https://deco.is.tue.mpg.de/	Dataset providing dense labels of human-object contact points to study body-part interaction in the wild
COCO Lin et al. (2014)	CC-BY 4.0	https://cocodataset.org/	Common Objects in Context benchmark for object detection, segmentation and captioning within rich scenes
FlickrSC3D Fieraru et al. (2021)	CC-BY (ages)	(im- https://ci3d.imar.ro/)	Flickr images annotated for self-contact body-part pairs projected onto 3D human meshes
LFW Huang et al. (2008)	Non-commercial research use	http://vis-www.cs.umass.edu/lfw/	Unconstrained face-photo collection for benchmarking face verification and identity clustering
EgoGesture Zhang et al. (2018)	Not distributed (Custom license)	https://github.com/KardinalZhu/EgoGesture	Egocentric RGB-D video dataset of dynamic hand gestures captured in varied environments
HaGRID Kapitanov et al. (2024)	CC BY-SA 4.0 / Custom (Variant of CC BY-SA 4.0)	https://github.com/hukenovs/hagrid	Image dataset of diverse hand gestures labeled for gesture detection and classification
BOLD Luo et al. (2020)	Publicly available	https://cydar.ist.psu.edu/emotionchallenge/dataset.php	Body-Language-in-the-Wild dataset aimed at recognizing bodily expressions of emotion in video
Emotic Kosti et al. (2017)	Used for commercial research and education purposes	https://github.com/rkosti/emotic	Contextual emotion-recognition dataset with images labeled by categorical and dimensional affect annotations
Ego4D Grauman et al. (2022)	Accessible for academic research (Custom License)	https://ego4d-data.org/	Egocentric video benchmark covering daily activities for tasks such as episodic memory, interaction and forecasting
Ego-Exo4D Grauman et al. (2024)	Accessible for academic research (Custom License)	https://ego-exo4d.github.io/	Dataset with synchronized first- and third-person videos plus narrations for skill and action analysis
HAA500 Chung et al. (2021)	MIT License	https://github.com/escalab/HAA500	Human-centric atomic-action dataset with fine-grained labels for action recognition
ActivityNet Fathi et al. (2015)	CC BY-NC-SA 4.0	http://activity-net.org/	Benchmark of untrimmed videos for temporal action localization and activity understanding
ActionAtlas Salehi et al. (2024)	Publicly available (meta-data)	https://github.com/mrsalehi/action-atlas	Sports-video QA benchmark that requires identifying subtle domain-specific actions

Continued on next page

Continuation of Table 1 (continued from previous page)

Name	License	URL	Overview
EgoExo Learn Huang et al. (2024)	MIT License	https://huggingface.co/datasets/hyf015/EgoExoLearn	Dataset where learners record egocentric videos while following exocentric demonstrations to study skill acquisition
HoloAssist Wang et al. (2023)	DLAv2 license	https://holoassist.github.io/	Egocentric mixed-reality dataset capturing remote human collaboration on physical manipulation tasks
NurViD Ming et al. (2023)	CC-BY (metadata)	https://github.com/minghu0830/NurViD-benchmark	Video benchmark of nursing procedures annotated for procedural steps and action recognition
ProBio Cui et al. (2023)	CC BY-NC-SA	https://probio-dataset.github.io/	Bioscience-lab dataset with hierarchical annotations for procedure tracking and action recognition
HA-ViD Zheng et al. (2023)	CC BY-NC 4.0	https://iai-hrc.github.io/ha-vid/	Industrial-assembly videos annotated for temporal and spatial understanding of human assembly actions
FineBio Yagi et al. (2024)	Limited to non-commercial research (Custom license)	https://sites.google.com/view/finebio	Dataset of biological experiments with hierarchical protocol-to-operation annotations for fine-grained action analysis
HiREST Zala et al. (2023)	MIT License	https://github.com/j-min/HiREST	Hierarchical video-moment dataset linking instructional video segments to retrieval queries and step captions
Captain Cook4D Peddi et al. (2024)	Apache license 2.0	https://captaincook4d.github.io/captain-cook/	CaptainCook4D provides egocentric cooking recordings enabling reasoning about procedural dependencies in recipes
LEMMA Jia et al. (2020)	Credentialed (gated) access	https://sites.google.com/view/lemma-egocentric-video/	Goal-directed multi-agent egocentric dataset with compositional atomic-action labels for complex activity understanding
Perception Test Pătrăucean et al. (2023)	CC-BY	https://github.com/deepmind/perception_test	DeepMind benchmark of real-world videos combining vision, audio and language tasks to probe perception and reasoning
EgoPlan-Bench Chen et al. (2023)	CC BY-NC-SA (EPIC-KITCHENS)& Custom License (Ego4D)	https://github.com/ChenYi99/EgoPlan	Benchmark evaluating multimodal large-language models on next-action planning in egocentric videos
PISC Li et al. (2017)	CC-BY 4.0	https://zenodo.org/record/1059155	Image dataset with person pairs annotated for social-relationship categories and roles
ViSR Liu et al. (2019)	Publicly available	https://xinchengliu.com/	Video dataset labeled with everyday social-relationship types between people
Crowd Activity Dataset Wang et al. (2022)	Publicly available	https://github.com/MCLAB-OCR/KnowledgeMiningWithSceneText/blob/main/README_datasets.md	Image collection categorized to distinguish crowd activities and interactions in diverse events

Continued on next page

Continuation of Table 1 (continued from previous page)

Name	License	URL	Overview
Struggle De-termination Dataset Feng et al. (2024)	Publicly available	https://github.com/FELIXFENG2019/Struggle-Determination	First-person video dataset assessing perceived struggle levels in assembly tasks
Crowd11 Dupont et al. (2017)	Publicly available (metadata)	https://paperswithcode.com/dataset/crowd11	Video dataset labeled with varied crowd-motion patterns for crowd-behaviour analysis
UCF-Crime Sultani et al. (2018)	For research purposes only.	https://www.crcv.ucf.edu/data/ucf-crime-dataset/	Surveillance video dataset with diverse anomaly categories for abnormal-event detection
CrisisMMD Alaoui et al. (2018)	CC BY-NC-SA 4.0	https://crisisnlp.qcri.org/crisismmd/	Multimodal Twitter dataset of disaster images and text annotated for crisis-response research

2 EXTRA DETAILS OF ANNOTATION SETUP

2.1 TASK-LEVEL ANNOTATION CONFIGURATION

Table 2: Annotation configuration for each task. Header “M.” stands for Modality, where I = Image, MI = Multi-Image, and V = Video.

Task	Source	M.	Explanation	Example
Spatial Relation	VSR	I	Perceive spatial relations among individuals and between individuals and objects.	Question: How would you describe the proximity of the person to the boat? Choices & Answer: A. in front of the boat B. behind the boat C. far from the boat D. on the boat E. next to the boat
Object tence	Exis- RefCOCOg	I	Identify the objects present in the image and determining the presence of specific target objects.	Question: What item appears in the picture? Choices & Answer: A. A green plate B. A green cup C. A green vase D. A red vase E. A blue vase
Human ence	Pres- RefCOCOg	I	Identify the individuals present in the image determining the presence of specific target individuals.	Question: Select the person that appears in the picture Choices & Answer: A. A man wearing a plaid dress shirt B. A man in a floral patterned shirt C. A man wearing a solid blue shirt D. A man wearing a striped red shirt E. A man in a plain white shirt
Body Posture	Motion-X	I	Recognize the fine-grained posture of the referenced person.	Question: What is the arrangement of the person’s body in the image, particularly regarding the state of specific body parts? Choices & Answer: A. Right thigh inclined, both legs at same height, left foot backward B. Both thighs vertical, right leg higher than left, left foot backward C. Right thigh vertical, left leg higher than right, left foot forward D. Both thighs vertical, left leg lower than right, left foot forward E. Right thigh horizontal, left leg higher than right, right foot forward

Continued on next page

Continuation of Table 2 (continued from previous page)

Task	Source	M.	Explanation	Example
Hand State	Motion-X	I	Identity the fine-grained hand states of the referenced person.	Question: How is the person's hand specifically arranged in the photograph? Choices & Answer: A. Both hands form a tight fist except for the thumbs, which are extended B. Fingers interlocked, thumbs pointing upward C. All fingers are closely positioned with both thumbs bent inward D. Thumbs wide apart, left fingers slightly bent, and right fingers slightly bent E. All fingers extended and close together, with thumbs sticking out to the sides
Hand-Object Interaction	HAKE-PaSta	I	Recognizing how the persons' hands interact with objects.	Question: How are the hands of the person shown in the image engaging with the sheep? Choices & Answer: A. Patting the sheep B. Feeding the sheep C. Pulling the sheep D. Carrying the sheep E. Pushing the sheep
Body Orientation	Orienta-MEBOW	I	Estimate the body orientation of the referenced person.	Question: What direction is this person's body turned towards in the picture? Choices & Answer: A. Directly backward B. Slightly to the left of forward C. Slightly to the right of forward D. Directly forward E. Directly to the left
Gaze Estimation	Estima-GazeFollow	I	Estimate the gaze direction of the referenced person.	Question: In the picture, which way is the seated woman's gaze most likely directed? Choices & Answer: A. Slightly upward and to the left B. Slightly downward and to the left C. Directly downward D. Slightly downward and to the right E. Directly to the left
Clothing Attribute	At-Clothing Attribute Dataset	I	Identifying clothing categories and attributes of the referenced person.	Question: What type of neckline design can be observed on the woman's dress, and what is the overall pattern on her clothing? Choices & Answer: A. Square neckline, plaid pattern B. V-shape neckline, plaid pattern C. V-shape neckline, floral pattern D. Round neckline, plaid pattern E. Scoop neckline, striped pattern
Accessory Recognition	ChaLearn LAP	I	Identifying the accessories of referenced person.	Question: What accessory pieces is the person in the image wearing? Choices & Answer: A. Earrings B. None C. A ring D. A necklace E. A watch
Bodypart Visibility	Visi-MPII Pose	I	Determine whether specific body-parts of the referenced individual are visible.	Question: Which upper body joints are completely visible in the image, specifically focusing on the shoulders, elbows, and wrists? Choices & Answer: A. Right elbow, right shoulder, thorax, left shoulder, left elbow B. Right shoulder, right elbow, left shoulder, left elbow, thorax C. Right wrist, right elbow, right shoulder, left shoulder, left elbow, left wrist D. Left shoulder, right shoulder, right wrist, left wrist E. Thorax, right wrist, right elbow, left wrist, left elbow

Continued on next page

Continuation of Table 2 (continued from previous page)

Task	Source	M.	Explanation	Example
Physical tribute	At- CelebA	I	Identify observable human physical attributes from images.	Question: Can you describe how the prominence of the person’s facial features might be balanced by the presence or absence of specific elements such as a beard or baldness? Choices & Answer: A. The presence of a thick beard and bushy eyebrows obscure other features like the big lips and receding hairline B. The absence of a receding hairline and presence of a small beard focuses attention away from the big nose and arching eyebrows C. The absence of a beard and baldness, coupled with prominent, curly hair and a receding hairline enhances the facial feature balance D. Minimal facial hair and a lack of baldness shift emphasis to the big nose, despite the styled, straight hair E. A full beard and a distinct presence of baldness shift focus towards their high cheekbones and narrow eyes
Human-Human Contact	FlickrCI3D (annotation, image from Flickr)	I	Identify specific body contact regions between two individuals.	Question: Can you identify any contact involving the third person from the right’s right hand back side fingers and the fourth person from the right? Choices & Answer: A. Yes, but only with the fourth person from the right’s left thigh. B. Yes, their right hand palms are in contact instead. C. Yes, the third person from the right’s fingers touch the fourth person from the right’s shoulder. D. Yes, the third person from the right’s right hand back side fingers are in contact with the fourth person from the right’s waist. E. No, there is no involvement with the fourth person from the right’s body.
Human-Object Contact	Hands23, DECO (annotation, image from Flickr), COCO (image)	I	Identify specific body contact regions between an object and an individual.	Question: What parts of this person’s body make contact with the cell phone? Choices & Answer: A. The right hand fingers (excluding palm) and the right ear. B. The left hand fingers (excluding palm) and the right ear. C. The right hand fingers (excluding palm) and the left ear. D. The left hand fingers (excluding palm) only. E. The left hand fingers (excluding palm) and the left ear.
Human Contact	Self- FlickrSC3D (annotation, image from Flickr)	I	Identify specific body contact parts of an individual.	Question: Is the ‘Left Elbow Pit’ of the girl wearing a red top in contact with their ‘Left Thigh’? Choices & Answer: A. Yes, the ‘Left Elbow Pit’ of the girl wearing a red top is in contact with their ‘Right Thigh’. B. Yes, the ‘Left Elbow Pit’ of the girl wearing a red top is in contact with their ‘Left Thigh’. C. No, the ‘Left Elbow Pit’ of the girl wearing a red top is not in contact with their ‘Left Thigh’. D. The ‘Right Elbow Pit’ of the girl wearing a red top is in contact with their ‘Left Thigh’. E. The ‘Left Elbow Pit’ of the girl wearing a red top is in contact with their ‘Left Forearm Outer Side’.
Face tion	Recogni- LFW	MI	Determine whether the faces in two images belong to the same individual.	Question: Is the person in the first picture the same as the person in the second picture? Choices & Answer: A. yes B. no

Continued on next page

Continuation of Table 2 (continued from previous page)

Task	Source	M.	Explanation	Example
Identity Clustering	LFW	MI	Identify which images depict the same individual among a set of images.	Question: In these 5 images, which ones are of the same person? Choices & Answer: A. The second, the fourth, and the fifth one are the same person. B. The second, the third, and the fourth one are the same person. C. The third, the fourth, and the fifth one are the same person. D. The third and the fifth one are the same person. E. The second, the third, and the fifth one are the same person.
Gesture	EgoGesture, Ha-GRID	I+V	Recognize human gestures and their meanings in images and videos.	Question: Which hand gesture is performed first, 'zoom in with fingers' or 'rotate fingers counterclockwise'? Choices & Answer: A. The gesture 'rotate fingers counterclockwise' does not appear. B. The gesture 'zoom in with fingers' does not appear. C. The gesture 'zoom in with fingers' appears first. D. Neither 'zoom in with fingers' nor 'rotate fingers counterclockwise' appears. E. The gesture 'rotate fingers counterclockwise' appears first.
Emotion	BOLD, Emotic	I+V	Perceive the emotional state of the referenced individual.	Question: What do this woman's posture and facial expressions reveal about their feelings? Choices & Answer: A. Engaged and entirely approving B. Engaged but slightly disapproving C. Surprised and unengaged D. Disinterested and neutral E. Completely disapproving and unengaged
Basic Action	Ego4d, Exo4D, HAA500, ActivityNet	Ego-V	Recognize basic human actions in video clips.	Question: What type of sport are the two individuals primarily engaged in throughout the video? Choices & Answer: A. Tennis is the primary activity, with shots of players performing serves and volleys indoors. B. The primary activity is volleyball, shown through indoor scenes featuring spikes and serves. C. Futsal is the primary sport as men demonstrate soccer techniques in an indoor stadium setting. D. Baseball is being played in an indoor stadium, showing pitching and catching techniques prominently. E. Basketball is their main focus, utilizing indoor courts to demonstrate dribbling and shooting skills.
Knowledge-Based Action	Ego-Exo4D, ActionAtlas(Used as-is), EgoExoLearn, HoloAssist, NurViD, ProBio, HA-ViD, FineBio	V	Recognize human action types within a specific domain.	Question: What best describes the move performed by the player who is farther from the camera? Choices & Answer: A. Backhand block B. Backhand drive C. Reverse pendulum serve D. Forehand loop E. Backhand flick

Continued on next page

Continuation of Table 2 (continued from previous page)				
Task	Source	M.	Explanation	Example
Sequential Action	Ego4D, HiREST	V	Recognize sequences of consecutive actions in video clips and their temporal relationships.	Question: What sequence of actions can be observed in the video? Please list them in order. Choices & Answer: A. The camera wearer raises his legs up, does sit-ups, raises his legs up again, does more sit-ups, folds his legs, raises his hands, looks at the watch, and then moves his hands in front. B. The camera wearer raises his legs up, raises his legs up again, folds his legs, raises his hands, moves his hands in front, and exercises on the mat. C. The camera wearer raises his legs up, does sit-ups, raises his legs up again, moves his foot, does sit-ups, folds his legs, puts the leg down, raises his hands, looks at the watch, and then moves his hands in front. D. The camera wearer does sit-ups, raises his legs up, folds his legs, raises his hands, looks at the watch, and then moves his hands in front. E. The camera wearer raises his legs up, raises his legs up again, does sit-ups, folds his legs, looks at the watch, and then moves his hands in front.
Procedure Dependence	CaptainCook4D	V	Determine whether a pair of actions in a video can be reordered, and identify the reason if they cannot.	Question: In the video, can the step 'Microwave until the cheese melts' and the step 'Replace the top of the English muffin' swap execution order? Choices & Answer: A. Yes, they can swap execution order. B. No, because although 'Replace the top of the English muffin' does not depend on 'Microwave until the cheese melts', its dependent predecessor actions had not been fully completed before 'Microwave until the cheese melts' was performed. C. No, because although 'Microwave until the cheese melts' does not depend on 'Replace the top of the English muffin', its dependent predecessor actions had not been fully completed before 'Replace the top of the English muffin' was performed. D. No, because 'Replace the top of the English muffin' depends on 'Microwave until the cheese melts'. E. No, because 'Microwave until the cheese melts' depends on 'Replace the top of the English muffin'.
Multiple Human Sequence Action	LEMMA	V		Question: What is the final location of both the cup and the waterpot at the end of the video? Choices & Answer: A. Both the cup and the waterpot are located on the table, visible to both another person and camera wearer. B. The cup is on the side of the waterdispenser and the waterpot sits on the sofa next to the camera wearer. C. The cup is held by another person, while the waterpot remains filled with water on the waterdispenser side. D. The cup is placed on the table and the waterpot is on the sofa. E. The cup ends up on the table, and the waterpot is put on the floor.

Continued on next page

Continuation of Table 2 (continued from previous page)

Task	Source	M.	Explanation	Example
Irrelevant Action	Ego4D (video, goal-step annotation), Perception Test (Used as-is)	V	Recognize sequential actions performed in parallel by two individuals, including their interactions and temporal relationships between actions.	Question: Considering the goal of the camera wearer reflected in this first-person video, which activities could be omitted to streamline the process? Choices & Answer: A. No activities unrelated to the overall goal are present in the video. B. Cook flatbread. C. Stir zucchini with the spatula. D. Transfer cooked flatbread to bowl. E. Mold dough with flour.
Goal Planning	EgoPlan-Bench (Used as-is)	V	Infer the next appropriate action based on the task goal and observed progress in the video.	Question: Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to transfer food from sieve to chopping board? Choices & Answer: A. put chopstick on chopping board B. put sieve in sink C. pour sieve on chopping board D. adjust bracelet E. pick up chopstick
Human Comparison	COCO, Crowd Activity Dataset	I	Identify similarities and differences between two individuals regarding attributes, positions, and roles.	Question: What do the two people on the far right in the image have in common? Choices & Answer: A. They both appear distressed. B. They both show two rows of teeth. C. They both wear a blue helmet. D. They both have sunglasses hanging in front of their clothes. E. They both have a green wristband on their left hand.
Social Relation	PISC, ViSR	I+V	Infer the social relationship between two referenced individuals.	Question: What potential social networks could the people in the scene belong to? Choices & Answer: A. Professional networking sites B. International travel clubs C. Local sports teams forums D. Family-oriented platforms E. Elderly community groups
Group Activity	COCO, Crowd Activity Dataset	I	Understand group activities, interactions between people, and roles of people in activities.	Question: What does the illustration in the picture say? Choices & Answer: A. The picture shows that Brexit has no impact. B. Brexit is the wrong choice. C. This is a neutral illustration, so it is hard to judge. D. Brexit was the right thing to do. E. Brexit is a brilliant decision as depicted in the illustration.
Struggle Detection	Struggle termination Dataset	V	Assess the level of struggle in first-person actions depicted in the video.	Question: Does the video show the person facing challenges while carrying out the task? Choices & Answer: A. Clear signs of struggle, like repeated attempts and jerky movements, showing uncertainty and lack of skill. B. Actions are mostly smooth, with occasional minor mistakes or adjustments. C. Signs of difficulty, such as errors or pauses, indicate some inexperience but effort to continue. D. Actions are executed smoothly and confidently with no visible difficulty.

Continued on next page

Continuation of Table 2 (continued from previous page)				
Task	Source	M.	Explanation	Example
Crowd Event	Crowd11, Crowd Activity Dataset	I+V	Understand individual trajectories and movements within crowds, as well as overall crowd dynamics.	Question: What is the movement trajectory of the person riding the motorcycle with a helmet? Choices & Answer: A. Standing near a store, looking at the products. B. Move from the right side of the frame to the left side. C. The person is getting closer to the camera. D. Move from the left side of the frame to the right side. E. The person is getting farther from the camera.
Cultural Event	ChaLearn LAP	I	Understand the content and places of cultural events in the images.	Question: What cultural celebration, involving fruit-throwing, is depicted in the image? Choices & Answer: A. Battle of the Oranges B. Grape Harvest Festival C. La Tomatina D. Apple Fest E. Lemon Festival
Crime Abnormal	UCF-Crime	V	Understand and explain the type, dynamics, and individual behaviors involved in abnormal activities.	Question: Did the people in the video enter the white car voluntarily? Choices & Answer: A. Yes, the people entered the white car voluntarily to avoid the confrontation. B. No, the people were already inside the car and did not exit at any point. C. ... D. ... E. No, the two men in black forcefully dragged out the people from the white car.
Disaster Under-standing	CrisisMMD	I	Classify disaster-related tweets and images by informativeness, humanitarian category, and damage severity.	Question: How would you describe the impact on the infrastructure observed in this area following the event? Choices & Answer: A. No visible damage to buildings or roads B. Mild structural damages with most buildings intact C. Buildings completely destroyed, debris everywhere D. Heavy flooding with multiple collapsed bridges E. Severe roof damage to numerous houses but no flooding

3 LEADERBOARD OF EACH TASK

The leaderboard of each task within perception and comprehension level is presented in Table 3, 4, 5.

Table 3: Leaderboards of models with top-8 performance for each task in Human-P and Human-C (1/3). “Rk” denotes the models’ ranks, and “Acc” denotes their accuracy.

(1) Human-P&Human-C			(2) Human-P			(3) Human-C		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	InternVL3-78B	62.77	1	InternVL3-38B	66.15	1	o4-mini	60.42
2	InternVL3-38B	62.74	2	InternVL3-78B	65.34	2	InternVL3-78B	60.20
3	o4-mini	62.28	3	Gemini-2.5-Flash	64.66	3	InternVL3-38B	59.32
4	InternVL2.5-78B	60.58	4	o4-mini	64.13	4	InternVL2.5-78B	58.21
5	Gemini-2.5-Flash	60.02	5	InternVL2.5-38B	63.07	5	InternVL2.5-38B	56.76
6	InternVL2.5-38B	59.92	6	LLaVA-OneVision-72B	62.96	6	Gemini-2.5-Flash	55.38
7	InternVL3-14B	58.96	7	InternVL2.5-78B	62.95	7	InternVL3-14B	55.11
8	LLaVA-OneVision-72B	57.98	8	InternVL3-14B	62.82	8	Qwen2.5-VL-72B	54.48
(4) Spatial Relation			(5) Object Existence			(6) Human Presence		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	InternVL2.5-78B	82.08	1	InternVL2.5-38B	88.00	1	InternVL3-78B	90.20
2	InternVL3-78B	81.50	2	InternVL3-78B	87.93	2	InternVL3-38B	88.24
3	Gemini-2.5-Flash	80.35	3	InternVL3-38B	87.36	3	InternVL2.5-78B	87.25
4	InternVL2.5-38B	78.74	4	LLaVA-OneVision-72B	86.21	4	InternVL2.5-38B	86.27
5	Qwen2.5-VL-72B	78.74	5	InternVL3-8B	85.63	5	LLaVA-OneVision-72B	85.78
6	InternVL3-14B	78.61	6	Oryx-1.5-32B	85.63	6	LLaVA-OneVision-7B	85.78
7	InternVL3-38B	78.61	7	InternVL2.5-78B	85.06	7	InternVL3-8B	85.29
8	o4-mini	77.46	8	InternVL2.5-4B	84.00	8	Qwen2.5-VL-72B	85.29
(7) Body Posture			(8) Hand State			(9) Hand-Object Interaction		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	Qwen2.5-VL-72B	61.47	1	Gemini-2.5-Flash	56.18	1	InternVL3-78B	58.73
2	InternVL2.5-78B	58.72	2	o4-mini	51.69	2	InternVL3-8B	58.41
3	InternVL3-78B	54.13	3	Gemini-2.0-Flash	51.69	3	o4-mini	57.78
4	LLaVA-OneVision-72B	54.13	4	Gemini-1.5-Pro	48.31	4	Gemini-2.0-Flash	56.83
5	Oryx-1.5-32B	53.01	5	InternVL3-38B	47.19	5	InternVL2.5-78B	56.51
6	InternVL3-38B	52.29	6	Qwen2.5-VL-72B	46.07	6	Gemini-2.5-Flash	56.19
7	LLaVA-Video-72B	52.29	7	InternVL3-14B	43.82	7	GPT-4o	55.56
8	InternVL2.5-38B	51.38	8	InternVL2.5-2B	42.70	8	InternVL2.5-38B	55.56
(10) Body Orientation			(11) Gaze Estimation			(12) Clothing Attribute		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	o4-mini	50.00	1	o4-mini	55.73	1	InternVL3-78B	86.89
2	InternVL3-14B	41.58	2	Gemini-2.0-Flash	55.73	2	LLaVA-OneVision-72B	83.61
3	InternVL3-38B	41.58	3	Gemini-1.5-Pro	52.67	3	LLaVA-Video-72B	83.61
4	Qwen2.5-VL-72B	38.74	4	InternVL3-38B	51.15	4	Gemini-2.5-Flash	82.79
5	Gemini-2.5-Flash	38.42	5	MiniCPM-o-2.6	50.38	5	InternVL3-38B	82.79
6	InternVL2.5-38B	38.22	6	Qwen2.5-VL-72B	50.38	6	InternVL2.5-78B	81.15
7	InternVL3-8B	36.84	7	Gemini-1.5-Flash	48.09	7	Doubao-1.5-vision-pro	80.33
8	InternVL3-78B	34.74	8	InternVL3-14B	48.09	8	o4-mini	78.69
(13) Accessory Recognition			(14) Bodypart Visibility			(15) Physical Attribute		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	o4-mini	94.26	1	o4-mini	48.44	1	InternVL3-78B	70.24
2	Gemini-2.5-Flash	94.26	2	Qwen2.5-VL-72B	44.53	2	LLaVA-OneVision-72B	70.24
3	InternVL3-78B	94.26	3	LLaVA-Video-72B	42.97	3	LLaVA-Video-72B	70.00
4	NVILA-8B	94.26	4	InternVL3-14B	42.19	4	InternVL3-38B	68.45
5	Gemini-2.0-Flash	92.62	5	InternVL3-78B	42.19	5	InternVL3-14B	67.26
6	InternVL2.5-38B	92.62	6	InternVL2.5-38B	39.84	6	o4-mini	66.07
7	InternVL2.5-4B	92.62	7	InternVL2.5-78B	39.84	7	InternVL2.5-78B	66.07
8	InternVL3-8B	92.62	8	InternVL3-8B	39.84	8	InternVL2.5-38B	65.29

Table 4: Leaderboards of models with top-8 performance for each task in Human-P and Human-C (2/3).

(16) Human-Object Contact			(17) Human-Human Contact			(18) Human Self-Contact		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	InternVL2.5-38B	70.45	1	Gemini-2.0-Flash	50.00	1	LLaVA-Video-72B	73.17
2	InternVL3-38B	67.42	2	Gemini-2.5-Flash	48.53	2	LLaVA-OneVision-72B	70.73
3	Gemini-2.5-Flash	66.67	3	InternVL2.5-78B	47.06	3	InternVL3-14B	62.20
4	InternVL3-78B	65.91	4	InternVL3-38B	47.06	4	InternVL3-38B	59.76
5	Claude-3.5-Sonnet-v2	64.39	5	Grok-2-Vision	45.59	5	Oryx-1.5-7B	51.22
6	LLaVA-OneVision-72B	64.39	6	Qwen2.5-VL-72B	44.78	6	NVILA-8B	50.00
7	InternVL2.5-78B	62.88	7	Doubao-1.5-vision-pro	44.12	7	Gemini-2.0-Flash	46.34
8	o4-mini	62.12	8	InternVL3-78B	42.65	8	InternVL3-78B	43.90
(19) Face Recognition			(20) Identity Clustering			(21) Gesture		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	Gemini-2.5-Flash	87.00	1	Gemini-2.5-Flash	100.00	1	o4-mini	63.87
2	Doubao-1.5-vision-pro	85.00	2	InternVL3-78B	96.84	2	InternVL3-38B	52.19
3	Gemini-2.0-Flash	83.00	3	o4-mini	95.79	3	Gemini-2.5-Flash	51.46
4	Claude-3.5-Sonnet-v2	82.00	4	InternVL3-38B	95.79	4	GPT-4o	51.09
5	Gemini-1.5-Pro	77.00	5	InternVL2.5-38B	91.58	5	Gemini-2.0-Flash	48.18
6	InternVL2.5-38B	77.00	6	InternVL3-14B	91.58	6	InternVL2.5-78B	47.81
7	o4-mini	76.00	7	InternVL2.5-78B	89.47	7	InternVL3-78B	47.81
8	InternVL3-38B	76.00	8	MiniCPM-V-2.6	78.95	8	InternVL2.5-38B	47.08
(22) Emotion			(23) Basic Action			(24) Knowledge-Based Action		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	o4-mini	53.69	1	InternVL3-38B	73.82	1	InternVL3-78B	60.46
2	MiniCPM-V-2.6	53.69	2	InternVL3-78B	72.82	2	InternVL3-38B	60.13
3	InternVL2.5-78B	53.00	3	Qwen2.5-VL-72B	70.57	3	InternVL2.5-78B	58.82
4	MiniCPM-o-2.6	53.00	4	InternVL2.5-78B	70.32	4	o4-mini	57.03
5	InternVL3-14B	52.71	5	o4-mini	69.83	5	InternVL2.5-38B	56.59
6	InternVL2.5-4B	52.22	6	Gemini-2.0-Flash	67.33	6	Qwen2.5-VL-72B	55.28
7	InternVL3-8B	52.22	7	Gemini-1.5-Pro	66.33	7	GPT-4o	53.92
8	LLaVA-Video-72B	51.23	8	InternVL3-14B	66.33	8	Gemini-2.0-Flash	52.52
(25) Sequential Action			(26) Goal Planning			(27) Procedure Dependence		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	InternVL2.5-38B	59.80	1	InternVL3-38B	42.08	1	InternVL3-38B	79.31
2	InternVL2.5-78B	58.82	2	InternVL3-78B	42.08	2	InternVL2.5-78B	78.16
3	Gemini-2.5-Flash	57.35	3	InternVL2.5-38B	32.05	3	Gemini-2.5-Flash	74.71
4	o4-mini	55.88	4	InternVL2.5-78B	31.66	4	InternVL3-78B	73.56
5	InternVL3-78B	54.90	5	InternVL3-14B	31.66	5	Oryx-1.5-32B	71.26
6	Qwen2.5-VL-72B	54.41	6	o4-mini	30.89	6	InternVL2.5-38B	70.11
7	Gemini-1.5-Pro	52.94	7	Qwen2.5-VL-72B	27.41	7	o4-mini	67.82
8	InternVL3-38B	52.94	8	InternVL3-2B	25.48	8	Claude-3.5-Sonnet-v2	62.07
(28) Multiple Human Sequence Action			(29) Irrelevant Action			(30) Human Comparison		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	InternVL2.5-38B	53.33	1	o4-mini	75.90	1	o4-mini	81.76
2	Doubao-1.5-vision-pro	50.00	2	Gemini-2.0-Flash	75.18	2	InternVL3-14B	76.47
3	InternVL3-14B	48.33	3	Gemini-1.5-Pro	68.71	3	InternVL3-38B	76.47
4	InternVL3-38B	45.00	4	Gemini-1.5-Flash	67.27	4	InternVL2.5-78B	74.71
5	o4-mini	41.67	5	InternVL2.5-38B	63.67	5	InternVL3-78B	73.53
6	InternVL2.5-78B	41.67	6	Claude-3.5-Sonnet-v2	63.31	6	InternVL2.5-38B	72.94
7	InternVL3-78B	41.67	7	Gemini-2.5-Flash	61.87	7	Gemini-2.5-Flash	72.35
8	Oryx-1.5-32B	41.67	8	Doubao-1.5-vision-pro	61.51	8	LLaVA-OneVision-72B	71.18

Table 5: Leaderboards of models with top-8 performance for each task in Human-P and Human-C (3/3).

(31) Social Relation			(32) Group Activity			(33) Struggle Detection		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	InternVL3-38B	65.33	1	o4-mini	82.51	1	InternVL2.5-2B	34.29
2	InternVL3-78B	63.82	2	InternVL3-78B	80.33	2	InternVL3-2B	32.86
3	InternVL3-14B	61.81	3	Gemini-2.5-Flash	79.78	3	Gemini-1.5-Pro	31.43
4	InternVL2.5-78B	61.31	4	LLaVA-OneVision-72B	79.78	4	InternVL2.5-1B	29.29
5	LLaVA-Video-72B	61.31	5	InternVL3-38B	79.23	5	GPT-4o	28.57
6	Claude-3.5-Sonnet-v2	60.80	6	InternVL2.5-78B	76.50	6	InternVL3-1B	28.57
7	Gemini-2.0-Flash	60.80	7	InternVL3-14B	76.50	7	LLaVA-Video-72B	26.43
8	o4-mini	60.30	8	NVILA-8B	74.86	8	Oryx-1.5-7B	25.71
(34) Crowd Event			(35) Cultural Event			(36) Crime Abnormal		
Rk	Model	Acc	Rk	Model	Acc	Rk	Model	Acc
1	InternVL3-38B	59.09	1	o4-mini	85.65	1	InternVL3-78B	77.84
2	InternVL3-78B	55.19	2	GPT-4o	85.22	2	Gemini-2.5-Flash	73.51
3	o4-mini	54.55	3	Gemini-2.0-Flash	84.35	3	Gemini-1.5-Pro	71.35
4	InternVL3-14B	53.90	4	Claude-3.5-Sonnet-v2	83.48	4	Gemini-2.0-Flash	70.27
5	InternVL2.5-78B	51.95	5	Gemini-2.5-Flash	83.48	5	o4-mini	68.11
6	Qwen2.5-VL-72B	51.30	6	InternVL3-78B	82.61	6	InternVL2.5-78B	66.49
7	InternVL2.5-38B	50.65	7	Qwen2.5-VL-72B	81.74	7	Qwen2.5-VL-72B	65.41
8	Gemini-2.5-Flash	48.05	8	InternVL2.5-78B	80.87	8	InternVL3-38B	64.86
(37) Disaster Understanding								
Rk	Model	Acc						
1	InternVL3-78B	65.19						
2	InternVL3-14B	62.22						
3	InternVL2.5-38B	61.48						
4	InternVL2.5-78B	61.48						
5	InternVL3-38B	59.26						
6	Qwen2.5-VL-72B	59.26						
7	InternVL3-8B	58.52						
8	GPT-4o	57.04						

4 EXAMPLES OF ANNOTATION FILES

```

{
  "id": "2444322301141a22070619141c153c46",
  "input_path": [
    "2444322301141a22070619141c153c46_1.jpg"
  ],
  "input_type": 0,
  "question": "How is the person's body arranged in the picture, focusing on the condition
    of specific body parts?",
  "answer": "The right knee is almost completely bent with horizontal calves",
  "distractors": [
    "The right knee is slightly bent",
    "Both knees are bent at right angles",
    "The left knee is slightly bent",
    "Neither of the knees is bent"
  ],
  "correct_option": "E",
  "question_prompt": "How is the person's body arranged in the picture, focusing on the
    condition of specific body parts?\nA. Both knees are bent at right angles\nB. The
    right knee is slightly bent\nC. Neither of the knees is bent\nD. The left knee is
    slightly bent\nE. The right knee is almost completely bent with horizontal calves",
  "task": "Body Posture",
  "dimension": "Posture",
  "level": "Perception"
}

```

Figure 1: Example annotation of the task Body Posture

```

{
  "id": "244432290f1e07211c1419043144",
  "input_path": [
    "244432290f1e07211c1419043144_1.jpg"
  ],
  "input_type": 0,
  "question": "What is the exact gesture of the person's hand in the picture?",
  "answer": "Both hands closed in tight fists",
  "distractors": [
    "Both hands making peace signs with all fingers fully extended",
    "Left index and both middles close to other fingers; both thumbs spread apart; right
    fingers barely bent",
    "Left hand is waving with fingers extended, right hand forming a cup shape",
    "Left hand with fingers forming an arch, right hand thumb up gesture"
  ],
  "correct_option": "C",
  "question_prompt": "What is the exact gesture of the person's hand in the picture?\nA.
    Left hand with fingers forming an arch, right hand thumb up gesture\nB. Both hands
    making peace signs with all fingers fully extended\nC. Both hands closed in tight
    fists\nD. Left hand is waving with fingers extended, right hand forming a cup
    shape\nE. Left index and both middles close to other fingers; both thumbs spread
    apart; right fingers barely bent",
  "task": "Hand State",
  "dimension": "Posture",
  "level": "Perception"
}

```

Figure 2: Example annotation of the task Hand State

```

{
  "id": "244432260f0a06371b01040c0f040a1d062a5e",
  "input_path": [
    "244432260f0a06371b01040c0f040a1d062a5e_1.jpg"
  ],
  "input_type": 0,
  "question": "What direction is the girl lying down in the picture most likely to be
    looking at?",
  "answer": "Slightly downward",
  "distractors": [
    "Directly downward",
    "Slightly upward",
    "Slightly to the right",
    "Slightly to the left"
  ],
  "correct_option": "C",
  "question_prompt": "What direction is the girl lying down in the picture most likely to be
    looking at?\nA. Slightly upward\nB. Directly downward\nC. Slightly downward\nD.
    Slightly to the right\nE. Slightly to the left",
  "task": "Gaze Estimation",
  "dimension": "Posture",
  "level": "Perception"
}

```

Figure 3: Example annotation of the task Gaze Estimation

```

{
  "id": "244432291b1d021c271707040d04201d06010c021a2f52",
  "input_path": [
    "244432291b1d021c271707040d04201d06010c021a2f52_1.jpg"
  ],
  "input_type": 0,
  "question": "What parts of this person's body make contact with the cell phone?",
  "answer": "The left hand fingers (excluding palm) and the left ear.",
  "distractors": [
    "The right hand fingers (excluding palm) and the right ear.",
    "The left hand fingers (excluding palm) and the right ear.",
    "The right hand fingers (excluding palm) and the left ear.",
    "The left hand fingers (excluding palm) only."
  ],
  "correct_option": "D",
  "question_prompt": "What parts of this person's body make contact with the cell phone?\nA.
    The right hand fingers (excluding palm) and the right ear.\nB. The left hand fingers
    (excluding palm) only.\nC. The right hand fingers (excluding palm) and the left
    ear.\nD. The left hand fingers (excluding palm) and the left ear.\nE. The left hand
    fingers (excluding palm) and the right ear.",
  "task": "Human-Object Contact",
  "dimension": "Contact",
  "level": "Understanding"
}

```

Figure 4: Example annotation of the task Human-Object Contact

```

{
  "id": "244432291b1d021c3b1001072d1f0d060916193e5c",
  "input_path": [
    "244432291b1d021c3b1001072d1f0d060916193e5c_1.png"
  ],
  "input_type": 0,
  "question": "Do the hands of the man wearing a yellow jacket touch each other excluding
    their fingers?",
  "answer": "Yes, the right and left hand palms, excluding fingers, of the man wearing a
    yellow jacket are in contact.",
  "distractors": [
    "No, the right and left hand palms, excluding fingers, of the man wearing a yellow
    jacket are not in contact.",
    "Yes, the right and left hand palms, including fingers, of the man wearing a yellow
    jacket are in contact.",
    "Only the right hand palm of the man wearing a yellow jacket, excluding fingers, is in
    contact with their left hand backside fingers.",
    "Only the left hand palm of the man wearing a yellow jacket, excluding fingers, is in
    contact with their right hand backside fingers."
  ],
  "correct_option": "B",
  "question_prompt": "Do the hands of the man wearing a yellow jacket touch each other
    excluding their fingers?\nA. Yes, the right and left hand palms, including fingers,
    of the man wearing a yellow jacket are in contact.\nB. Yes, the right and left hand
    palms, excluding fingers, of the man wearing a yellow jacket are in contact.\nC. Only
    the right hand palm of the man wearing a yellow jacket, excluding fingers, is in
    contact with their left hand backside fingers.\nD. Only the left hand palm of the man
    wearing a yellow jacket, excluding fingers, is in contact with their right hand
    backside fingers.\nE. No, the right and left hand palms, excluding fingers, of the
    man wearing a yellow jacket are not in contact.",
  "task": "Human Self-Contact",
  "dimension": "Contact",
  "level": "Understanding"
}

```

Figure 5: Example annotation of the task Human Self-Contact

```

{
  "id": "244432280a150d0601011422020510060d07040f092f51",
  "input_path": [
    "244432280a150d0601011422020510060d07040f092f51_1.jpg",
    "244432280a150d0601011422020510060d07040f092f51_2.jpg",
    "244432280a150d0601011422020510060d07040f092f51_3.jpg",
    "244432280a150d0601011422020510060d07040f092f51_4.jpg",
    "244432280a150d0601011422020510060d07040f092f51_5.jpg",
    "244432280a150d0601011422020510060d07040f092f51_6.jpg"
  ],
  "input_type": 1,
  "question": "In these 6 images, which ones are of the same person?",
  "answer": "The third, the fifth, and the sixth one are the same person.",
  "distractors": [
    "The fourth and the sixth one are the same person.",
    "The second and the third one are the same person; the fourth and the fifth one are the
    same person.",
    "The second and the fourth one are the same person.",
    "The first and the fourth one are the same person."
  ],
  "correct_option": "A",
  "question_prompt": "In these 6 images, which ones are of the same person?\nA. The third,
    the fifth, and the sixth one are the same person.\nB. The first and the fourth one
    are the same person.\nC. The second and the fourth one are the same person.\nD. The
    fourth and the sixth one are the same person.\nE. The second and the third one are
    the same person; the fourth and the fifth one are the same person.",
  "task": "Identity Clustering",
  "dimension": "Identity",
  "level": "Understanding"
}

```

Figure 6: Example annotation of the task Identity Clustering

```

{
  "id": "2447322a001f141e0d110a042c1110170c340e15071f0d2d5a",
  "input_path": [
    "2447322a001f141e0d110a042c1110170c340e15071f0d2d5a_1.mp4"
  ],
  "input_type": 2,
  "question": "What best describes the action performed by the biker?",
  "answer": "Barspin",
  "distractors": [
    "Decade Air",
    "Flair barspin",
    "Double barspin",
    "Tailwhip"
  ],
  "correct_option": "A",
  "question_prompt": "What best describes the action performed by the biker?\nA. Barspin\nB.
    Decade Air\nC. Flair barspin\nD. Tailwhip\nE. Double barspin",
  "task": "Knowledge-Based Action",
  "dimension": "Behavior",
  "level": "Understanding"
}

```

Figure 7: Example annotation of the task Knowledge-Based Action

```

{
  "id": "2447322601110f220414030f071e042d5c",
  "input_path": [
    "2447322601110f220414030f071e042d5c_1.mp4"
  ],
  "input_type": 2,
  "question": "Considering the progress shown in the video and my current observation in the
    last frame, what action should I take next in order to dry and put away utensils?",
  "answer": "close tap",
  "distractors": [
    "open tap",
    "pick up towel",
    "hang towel on fridge",
    "rinse utensils"
  ],
  "correct_option": "C",
  "question_prompt": "Considering the progress shown in the video and my current observation
    in the last frame, what action should I take next in order to dry and put away
    utensils?\nA. open tap\nB. pick up towel\nC. close tap\nD. hang towel on fridge\nE.
    rinse utensils",
  "task": "Goal Planning",
  "dimension": "Procedure",
  "level": "Understanding"
}

```

Figure 8: Example annotation of the task Goal Planning

```

{
  "id": "2447322c1b1c171b181908291b1d021c3b101c140b1e001729161908011e3c40",
  "input_path": [
    "2447322c1b1c171b181908291b1d021c3b101c140b1e001729161908011e3c40_1.mp4"
  ],
  "input_type": 2,
  "question": "Does any object experience a change of state more than once during the
sequence of interactions between the camera wearer and another person?",
  "answer": "The cup experiences multiple states; it moves from shelf to table, table to
microwave, and finally microwave back to table.",
  "distractors": [
    "The milk undergoes changes of state from being in the fridge to being in the cup and
finally back into the fridge.",
    "The microwave door is opened and closed multiple times by both the camera wearer and
another person during the sequence.",
    "The milk changes state from cold when poured out to possibly warm inside the cup in
the end of the video.",
    "The camera wearer and another person handed over the cup to each other multiple times,
altering its state change."
  ],
  "correct_option": "E",
  "question_prompt": "Does any object experience a change of state more than once during the
sequence of interactions between the camera wearer and another person?\nA. The camera
wearer and another person handed over the cup to each other multiple times, altering
its state change.\nB. The milk changes state from cold when poured out to possibly
warm inside the cup in the end of the video.\nC. The microwave door is opened and
closed multiple times by both the camera wearer and another person during the
sequence.\nD. The milk undergoes changes of state from being in the fridge to being
in the cup and finally back into the fridge.\nE. The cup experiences multiple states;
it moves from shelf to table, table to microwave, and finally microwave back to
table.",
  "task": "Multiple Human Sequential Action",
  "dimension": "Procedure",
  "level": "Understanding"
}

```

Figure 9: Example annotation of the task Multiple Human Sequential Action

```

{
  "id": "244732291b1d021c2b1a00110f020a01071b3250",
  "input_path": [
    "244732291b1d021c2b1a00110f020a01071b3250_1.png"
  ],
  "input_type": 0,
  "question": "What do the two people on the far right in the image have in common?",
  "answer": "They both have a green wristband on their left hand.",
  "distractors": [
    "They both show two rows of teeth.",
    "They both appear distressed.",
    "They both have sunglasses hanging in front of their clothes.",
    "They both wear a blue helmet."
  ],
  "correct_option": "E",
  "question_prompt": "What do the two people on the far right in the image have in
common?\nA. They both appear distressed.\nB. They both show two rows of teeth.\nC.
They both wear a blue helmet.\nD. They both have sunglasses hanging in front of their
clothes.\nE. They both have a green wristband on their left hand.",
  "task": "Human Comparison",
  "dimension": "Relation",
  "level": "Understanding"
}

```

Figure 10: Example annotation of the task Human Comparison

```

{
  "id": "244732261c1f1602291619081819170b3743",
  "input_path": [
    "244732261c1f1602291619081819170b3743_1.jpg"
  ],
  "input_type": 0,
  "question": "Who is most likely to answer a reporter's question?",
  "answer": "The man in the red tie",
  "distractors": [
    "A man in a black and white striped tie",
    "A woman in a white blouse",
    "The man on the far right.",
    "The woman on the far right."
  ],
  "correct_option": "C",
  "question_prompt": "Who is most likely to answer a reporter's question?\nA. A man in a
    black and white striped tie\nB. The woman on the far right.\nC. The man in the red
    tie\nD. A woman in a white blouse\nE. The man on the far right.",
  "task": "Group Activity",
  "dimension": "Relation",
  "level": "Understanding"
}

```

Figure 11: Example annotation of the task Group Activity

```

{
  "id": "244732221c1f14162d03080f1a2f56",
  "input_path": [
    "244732221c1f14162d03080f1a2f56_1.mp4"
  ],
  "input_type": 2,
  "question": "What is the movement trajectory of the girl taking photos of others in the
    center of the frame at the beginning of the video?",
  "answer": "She initially remains mostly stationary, then starts moving toward the person
    being photographed.",
  "distractors": [
    "She remains mostly stationary throughout the video until it ends.",
    "She moves toward the person being photographed right from the beginning.",
    "Move from the center of the frame to the right side.",
    "Move from the center of the frame to the left side."
  ],
  "correct_option": "E",
  "question_prompt": "What is the movement trajectory of the girl taking photos of others in
    the center of the frame at the beginning of the video?\nA. Move from the center of
    the frame to the right side.\nB. She remains mostly stationary throughout the video
    until it ends.\nC. She moves toward the person being photographed right from the
    beginning.\nD. Move from the center of the frame to the left side.\nE. She initially
    remains mostly stationary, then starts moving toward the person being photographed.",
  "task": "Crowd Event",
  "dimension": "Scene",
  "level": "Understanding"
}

```

Figure 12: Example annotation of the task Crowd Event

```

{
  "id": "244732221c190e172917030e1c1d021e3744",
  "input_path": [
    "244732221c190e172917030e1c1d021e3744_1.mp4"
  ],
  "input_type": 2,
  "question": "What kind of unusual activity is depicted in the video?",
  "answer": "Fighting",
  "distractors": [
    "Assault",
    "Robbery",
    "Abuse",
    "Burglary"
  ],
  "correct_option": "D",
  "question_prompt": "What kind of unusual activity is depicted in the video?\nA.
    Burglary\nB. Robbery\nC. Assault\nD. Fighting\nE. Abuse",
  "task": "Crime Anomaly",
  "dimension": "Scene",
  "level": "Understanding"
}

```

Figure 13: Example annotation of the task Crime Anomaly

```

{
  "id": "5b4359",
  "input_path": [
    "5b4359_1.mp4"
  ],
  "input_type": 2,
  "question": "What is the reason for the protagonist repeatedly rubbing the paint surface
    with his hands in the video?",
  "cot": "The reasoning is based on visual cues from the video that introduce the assembled
    airplane, the Lockheed P-38J Lightning, which is known to be a World War II-era
    aircraft. Observations in the video show that final aesthetic details, such as
    chipped paint on the propellers, nose, and fuel tank, are intentionally created,
    indicating the desire to achieve an aged and weathered look. This conclusion is drawn
    by correlating the historical context of the aircraft with the deliberate technique
    of rubbing off the paint.",
  "answer": "He is intentionally removing paint to create an aged, weathered appearance on
    the aircraft.",
  "question_type": "Cause Reasoning",
  "domain": "Electric & Crafting",
  "level": "Reasoning",
  "meta_data": {
    "source": "YouTube",
    "url": "https://www.youtube.com/watch?v=xDwEATd3Whs",
    "interval": [
      0,
      1469
    ]
  }
}

```

Figure 14: Example annotation of Human-R: 1


```

{
  "id": "5b4c",
  "input_path": [
    "5b4c_1.mp4"
  ],
  "input_type": 2,
  "question": "Why is the man playing a video on his phone while busy using a flashlight? Did he happen to stumble upon this video? Answer \"yes\" or \"no\" and briefly explain the reason.",
  "cot": "The video displays textual cues indicating that the man is on a rescue mission to save a kitten trapped in the house's foundation, as he directs a flashlight toward a hole where the kitten is likely hiding. Visual focus shifts to a phone playing a video of a cat, complete with a visual of its open mouth suggesting it is making a sound, and the phone's screen later reveals a search record for 'cat talking.' This evidence, combined with the knowledge that cats are attracted to the sounds of their own species, supports the interpretation that the man intentionally searched for and used these specific videos to lure the kitten, rather than having discovered the video by chance.",
  "answer": "No, he did not stumble upon the video accidentally; he intentionally played videos with cat sounds to attract the kitten for rescue.",
  "question_type": "Prediction",
  "domain": "Household",
  "level": "Reasoning",
  "meta_data": {
    "source": "YouTube",
    "url": "https://www.youtube.com/watch?v=ygc7vJaVgJA",
    "interval": [
      0,
      135
    ]
  }
}

```

Figure 15: Example annotation of Human-R: 2

```

{
  "id": "5a4154",
  "input_path": [
    "5a4154_1.mp4"
  ],
  "input_type": 2,
  "question": "Why does the chef cut it the ingredients this way?",
  "cot": "The video shows the chef deliberately making a cross-shaped incision on the flat surface of the chestnut; by later observing her peeling and consuming the chestnut, it becomes clear that this method makes the shell easier to remove. Additionally, the chef's technique addresses the fact that chestnuts have internal moisture, which could lead to explosions when heated if there is no cut to vent the steam, thereby ensuring safety during the heating process.",
  "answer": "The chef cuts it this way to facilitate shell removal and prevent the chestnut from exploding during heating.",
  "question_type": "Cause Reasoning",
  "domain": "Household",
  "level": "Reasoning"
}

```

Figure 16: Example annotation of Human-R: 3

5 EXAMPLES OF EACH TASK IN HUMAN-P&C

Perception - Spatiality - Spatial Relation

Question:

Considering the arrangement of the motorcycle and the parking meter, where would you find the motorcycle in relation to the parking meter?

Options:

- (A) The motorcycle is in front of the parking meter.
- (B) The motorcycle is on the right of the parking meter.
- (C) The motorcycle is behind the parking meter.
- (D) The motorcycle is across the street from the parking meter.
- (E) The motorcycle is above the parking meter.



Answer: C

Figure 17: An example of the task “Spatial Relation”.

Perception - Spatiality - Object Existence

Question:

What type of vehicle is depicted in the image?

Options:

- (A) A modern white airplane with blue lettering.
- (B) An old dark green airplane with no lettering.
- (C) A blue helicopter with white lettering.
- (D) An old dark blue airplane with white lettering.
- (E) A dark blue submarine with white numbers.



Answer: D

Figure 18: An example of the task “Object Existence”.

Perception - Spatiality - Human Presence

Question:

Which of these individuals is depicted in the picture?

Options:

- (A) A man in a blue shirt and black shorts riding a bicycle and carrying a backpack.
- (B) A man in a white shirt and khaki shorts riding a bicycle and pulling a cart.
- (C) A man in a red shirt and khaki shorts walking alongside a bicycle.
- (D) A man in a white shirt and blue shorts skateboarding.
- (E) A man in a yellow shirt and khaki shorts riding a scooter.



Answer: B

Figure 19: An example of the task “Human Presence”.

Perception - Posture - Body Posture

Question:

How is the person's body configured in the image, with particular attention to the state of different body parts?

Options:

- (A) Both hands are shoulder-width apart, the knees are slightly bent, and the right elbow is partially bent while the left elbow is nearly bent.
- (B) Both hands are together, knees are fully bent, and the right elbow is straight while the left elbow is fully bent.
- (C) Both hands are stretched wide apart, the right knee is unbent, and both elbows are equally bent.
- (D) Both hands are close to the body, both knees are slightly bent, and both elbows are equally straight.
- (E) Both hands are shoulder-width apart, knees are straight, and both elbows are partially bent.



Answer: A

Figure 20: An example of the task “Body Posture”.

Perception - Posture - Hand State

Question:

How does the person's hand appear in the photograph?

Options:

- (A) The rings are closely positioned, thumbs are close together, and all fingers are fully extended.
- (B) Each finger including the rings and thumbs are evenly spaced and slightly bent.
- (C) The rings are tightly held together, thumbs are spread, and the right fingers are bent significantly.
- (D) All fingers, including thumbs and rings, are spread widely and distinctly bent.
- (E) The rings are slightly spread to other fingers, thumbs are far apart, and the left fingers are somewhat bent.



Answer: E

Figure 21: An example of the task “Hand State”.

Perception - Posture - Hand-Object Interaction

Question:

How does the individual in the photo use their hands to interact with the baseball_glove?

Options:

- (A) Holding the baseball glove up.
- (B) Throwing the baseball glove.
- (C) Catching the baseball glove.
- (D) Wearing the baseball glove on their hand.
- (E) Tying the laces of the baseball glove.



Answer: D

Figure 22: An example of the task “Hand-Object Interaction”.

Perception - Posture - Body Orientation

Question:

What is the body's orientation of the person wearing jeans in the context of the picture?

Options:

- (A) Facing directly backwards.
- (B) Facing slightly to the left from backview.
- (C) Fully facing to the left.
- (D) Facing forwards.
- (E) Facing slightly to the right from backview.



Answer: E

Figure 23: An example of the task “Body Orientation”.

Perception - Posture - Gaze Estimation

Question:

In the image, which direction is the man being pulled along for skiing by a man wearing the orange hat most likely to be facing with their eyes?

Options:

- (A) Slightly upward.
- (B) Directly rightwards.
- (C) Slightly downward.
- (D) Directly leftwards.
- (E) Straight downward.



Answer: C

Figure 24: An example of the task “Gaze Estimation”.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

Perception - Appearance - Clothing Attribute

Question:
Can you describe any specific design features and the general colorfulness of the dress worn by the female with low skin exposure?

Options:
(A) Round neckline and solid color.
(B) V-neckline and very colorful.
(C) Square neckline and very colorful.
(D) Round neckline and very colorful.
(E) Round neckline and two-tone.



Answer: D

Figure 25: An example of the task “Clothing Attribute”.

Perception - Appearance - Accessory Recognition

Question:
Which accessories can be seen on the person in the photo?

Options:
(A) Necklace, Bracelet.
(B) Necklace, Headphone.
(C) Earrings, Headband.
(D) Necklace, Hat.
(E) Scarf, Headband.



Answer: B

Figure 26: An example of the task “Accessory Recognition”.

Perception - Appearance - Bodypart Visibility

Question:

What might be the reasons for the asymmetrical visibility of the joints in the legs and arms of the person displayed?

Options:

- (A) The person could be positioned with one side closer to the camera, making joints on the other side less visible.
- (B) The person is wearing a uniform that covers all but the left side of their body.
- (C) The person is standing in a shadow that obscures the joints on their right side completely.
- (D) The image quality is poor, blurring out details on the right side of the body.
- (E) The light source is positioned to only illuminate the person's left side efficiently.



Answer: A

Figure 27: An example of the task "Bodypart Visibility".

Perception - Appearance - Physical Attribute

Question:

Can you describe the facial expression and hairstyle details that reflect this person's demeanor and hair color?

Options:

- (A) The person has arched eyebrows, bangs, blond wavy hair, big lips, and high cheekbones.
- (B) The person has bushy eyebrows, straight blond hair, and a serious expression.
- (C) The person has straight brown hair, low cheekbones, and narrow eyes.
- (D) The person has curly black hair, round face, and big lips.
- (E) The person has arched eyebrows, gray straight hair, and a youthful look.



Answer: A

Figure 28: An example of the task "Physical Attribute".

Perception - Contact - Human-Object Contact

Question:

What are the types of grasp and object interactions observed for the person's hands?

Options:

- (A) His right hand is using his fingertips to touch an object without grasping it. His left hand is not visible.
- (B) His left hand is touching an object without holding it. His right hand is not visible.
- (C) His left hand is touching an object without holding it, while his right hand has a circular precision grasp on a container.
- (D) His left hand is touching an object without holding it, while his right hand has a circular precision grasp on an object.
- (E) His left hand is touching an object without holding it, while his right hand is using his fingertips to touch an object without grasping it.



Answer: E

Figure 29: An example of the task “Human-Object Contact”.

Perception - Contact - Human-Human Contact

Question:

Are there any parts of the baby in the middle's left hand fingers in contact with the baby on the right?

Options:

- (A) No, the baby in the middle's left hand palm side fingers are not in contact with the baby on the right's right hand.
- (B) Yes, the baby in the middle's left hand palm side fingers are in contact with the baby on the right's right hand back side fingers.
- (C) Yes, the baby in the middle's left hand back side fingers are in contact with the baby on the right's right hand back side fingers.
- (D) Yes, the baby in the middle's left hand palm side fingers are in contact with the baby on the right's right hand palm.
- (E) No, the baby in the middle's left hand fingers only touch the baby on the right's upper abdomen.



Answer: B

Figure 30: An example of the task “Human-Human Contact”.

Perception - Contact - Human Self-Contact

Question:

Can you tell which parts of the woman donning a clown mask's body are touching each other?

Options:

- (A) The right ankle of the woman donning a clown mask is in contact with their right lower leg back side.
- (B) The left ankle of the woman donning a clown mask is in contact with their left lower leg back side.
- (C) The left ankle of the woman donning a clown mask is not in contact with their right lower leg back side.
- (D) The left ankle of the woman donning a clown mask is in contact with their right lower leg back side.
- (E) Both ankles of the woman donning a clown mask are in contact with their right lower leg back side.



Answer: D

Figure 31: An example of the task “Human Self-Contact”.

Perception - Identity - Face Recognition

Question:

Is the person in the first picture the same as the person in the second picture?

Options:

- (A) Yes.
- (B) No.



Answer: B

Figure 32: An example of the task “Face Recognition”.

Perception - Identity - Identity Clustering

Question:

In these 4 images, which ones are of the same person?

Options:

- (A) The third and the fourth one are the same person; the first and the second one are the same person.
 (B) The first, the second, and the third one are the same person.
 (C) The second and the third one are the same person.
 (D) The third and the fourth one are the same person.
 (E) The first and the second one are the same person; the third and the fourth one are the same person.



Answer: B

Figure 33: An example of the task “Identity Clustering”.

Comprehension - Behavior - Gesture

Question: Which hand gestures are performed in the video? Answer by listing them in the order they occur.

Options:

- (A) Gather fingers, thumb toward left, beckon, rotate fists counterclockwise, and static fist. (B) Static fist, thumbs forward, beckon, rotate fists counterclockwise, and gather fingers. (C) Static fist, thumb toward left, beckon, rotate fists counterclockwise, and gather fingers. (D) Beckon, thumb toward left, static fist, rotate fists counterclockwise, and gather fingers. (E) Static fist, thumb toward right, beckon, rotate fists counterclockwise, and gather fingers.



Answer: E

Figure 34: An example of the task “Gesture”.

Comprehension - Behavior - Emotion

Question: Which emotions best match the current state of the person with one arm over another's shoulder in the video?

Options:

- (A) He expresses surprise with wide eyes and raised eyebrows. (B) He looks worried with furrowed brows and a frown. (C) He reveals sadness, with a downturned mouth and drooping shoulders. (D) He shows anger through clenched jaws and furrowed brows. (E) He shows expressions of happiness mixed with disconnection.



Answer: E

Figure 35: An example of the task “Emotion”.

Comprehension - Behavior - Basic Action

Question: How do the participants utilize their body parts to maintain movement or balance during the main action, and what does this reveal about their coordination skills?

Options:

(A) Participants use their arms to catch while simultaneously jumping, revealing their agility and strong eyehand coordination. (B) They balance a ball on one hand while moving, indicating high skill in dynamic equilibrium management. (C) Body leaning and arm flailing helps them stay upright on the play structure, indicating moderate balancing ability. (D) Participants use their legs to pump and their hands to grip, demonstrating rhythmic and effective coordination skills. (E) Coordinated spinning with arms extended maintains centrifugal force, showing advanced control in rotational motion.



Answer: D

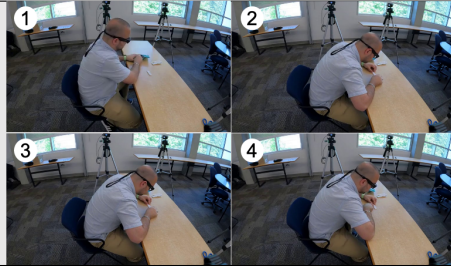
Figure 36: An example of the task “Basic Action”.

Comprehension - Behavior - Knowledge-Based Action

Question: What action is depicted in this video clip?

Options:

(A) Place the COVID-19 test kit on the counter. (B) Verify the serial number on the test kit. (C) Check the testing plate for the result. (D) Gather supplies needed for testing. (E) Review the test kit expiry date.



Answer: C

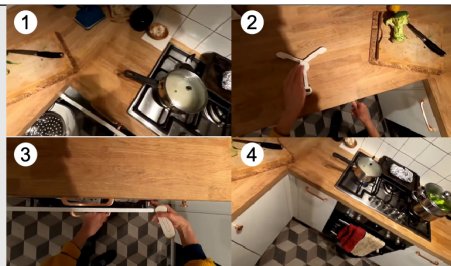
Figure 37: An example of the task “Knowledge-Based Action”.

Comprehension - Procedure - Sequential Action

Question: What sequence of actions can be observed in the video? Please list them in order.

Options:

(A) Moves the tray in the cupboard, places the scale in the cupboard, checks the label in the bin, adjust the pot of rice, place broccoli strainer in pot, and adjust the pot of broccoli. (B) Moves the tray in the cupboard, places the scale in the cupboard, place pan on scale, adjust the pot of rice, and adjust the pot of broccoli. (C) Moves the tray in the cupboard, places the scale in the cupboard, add rice in pot to measure, place pan on scale, adjust the pot of rice, adjust the pot of broccoli, and wash broccoli. (D) Moves the tray in the cupboard, places the scale in the cupboard, adjust the pot of rice, and adjust the pot of broccoli. (E) Moves the tray in the cupboard, places the scale in the cupboard, place pan on scale, adjust the pot of rice, and places the container in the cupboard.



Answer: D

Figure 38: An example of the task “Sequential Action”.

Comprehension - Procedure - Goal Planning

Question: Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to add butter paste to the stew and stir?

Options:

(A) take chopsticks. (B) put down chopsticks. (C) take spoon. (D) remove butter from spoon with chopsticks. (E) stir butter.



Answer: D

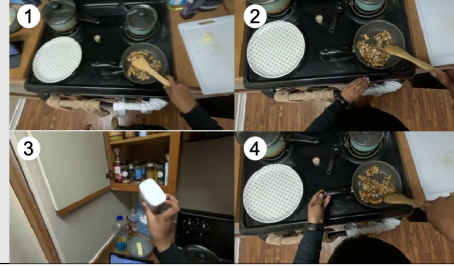
Figure 39: An example of the task “Goal Planning”.

Comprehension - Procedure - Procedure Dependence

Question: In the video, can the step 'Add minced garlic to the pan' and the step 'Season the pan with salt' swap execution order?

Options:

(A) No, because 'Add minced garlic to the pan' depends on 'Season the pan with salt'. (B) Yes, they can swap execution order. (C) No, because 'Season the pan with salt' depends on 'Add minced garlic to the pan'. (D) No, because although 'Season the pan with salt' does not depend on 'Add minced garlic to the pan', its dependent predecessor actions had not been fully completed before 'Add minced garlic to the pan' was performed. (E) No, because although 'Add minced garlic to the pan' does not depend on 'Season the pan with salt', its dependent predecessor actions had not been fully completed before 'Season the pan with salt' was performed.



Answer: C

Figure 40: An example of the task “Procedure Dependence”.

Comprehension - Procedure - Multiple Human Sequence Action

Question: What object does the camera wearer gesture towards and does this action precede or follow another person's utilization of the vacuum?

Options:

(A) The camera wearer gestures towards the vacuum and then the floor; both gestures come before and after another person's picking up the vacuum respectively. (B) The camera wearer gestures toward the sofa, and this happens before the vacuum is mentioned. (C) The floor is pointed at by the camera wearer after another person has finished using the vacuum. (D) The camera wearer points at the floor before another person starts to sweep the floor with it. (E) The camera wearer's pointing action towards the table precedes the deployment of the vacuum by another person.



Answer: A

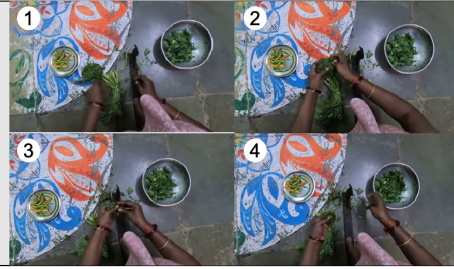
Figure 41: An example of the task “Multiple Human Sequence Action”.

Comprehension - Procedure - Irrelevant Action

Question: Considering the goal of the camera wearer reflected in this first-person video, which activities could be omitted to streamline the process?

Options:

(A) Slice fenugreek. (B) Chop spinach with knife. (C) Remove rope from spinach. (D) Gather vegetables in bowl. (E) No activities unrelated to the overall goal are present in the video.



Answer: D

Figure 42: An example of the task “Irrelevant Action”.

Comprehension - Relation - Human Comparison

Question:

In what aspect do the two people in the image differ?

Options:

(A) The person closer to the camera is wearing shoes, while the person farther away is barefoot.
 (B) The person closer to the camera is looking down, while the person farther away is looking up.
 (C) The person closer to the camera is using the parallel bars, while the person farther from the camera is helping the person closer practice.
 (D) The person closer to the camera is noticeably taller than the person farther away.
 (E) The person who are closer to the camera is taking photos, while the person who are farther away to the camera is skateboarding.



Answer: A

Figure 43: An example of the task “Human Comparison”.

Comprehension - Relation - Social Relation

Question: What social affiliations might exist among the individuals in the scene?

Options:

(A) Organizational hierarchy. (B) Book club membership. (C) Yoga class participants. (D) Neighborhood association. (E) University alumni.



Answer: A

Figure 44: An example of the task “Social Relation”.

Comprehension - Relation - Group Activity

Question:

What color sportswear is the player in the picture wearing who won the match?

Options:

- (A) Red.
- (B) Yellow.
- (C) Black.
- (D) Red and Black.
- (E) No Correct Option.



Answer: A

Figure 45: An example of the task “Group Activity”.

Comprehension - Relation - Struggle Detection

Question: Is the person in the video encountering obstacles or difficulty while performing the operation?

Options:

- (A) Actions are executed smoothly and confidently with no visible difficulty. (B) Actions are mostly smooth, with occasional minor mistakes or adjustments. (C) Signs of difficulty, such as errors or pauses, indicate some inexperience but effort to continue. (D) Clear signs of struggle, like repeated attempts and jerky movements, showing uncertainty and lack of skill.



Answer: C

Figure 46: An example of the task “Struggle Detection”.

Comprehension - Scene - Crowd Event

Question: What is the movement trajectory of the woman who appears on the right at the beginning of the video, carrying a pink bag and a beige handbag?

Options:

- (A) First, she moves to the right side of the frame, then after looking around, she moves downward in the frame. (B) She moves straight to the right side of the frame. (C) First, she looks around, then moves to the left side of the frame. (D) She remains in place, looking around. (E) No correct option.



Answer: A

Figure 47: An example of the task “Crowd Event”.

Comprehension - Scene - Cultural Event

Question:
What is the cultural significance of the event in the picture, taking place in Tunisia?

Options:
(A) It commemorates the annual harvest season.
(B) It celebrates desert life and Bedouin traditions.
(C) It marks the historical trade routes through the desert.
(D) It is a religious ceremony honoring local saints.
(E) It signifies the beginning of the fishing season in coastal areas.



Answer: B

Figure 48: An example of the task “Cultural Event”.

Comprehension - Scene - Crime Abnormal

Question: How many adults chased after the man following the incident with the child's mobile phone?

Options:
(A) Only the woman chased after the man, and no one else joined her.
(B) Three people went out to chase the man after he snatched the child's mobile phone.
(C) Two people, including the woman, went out to chase the man.
(D) Four individuals, reacting quickly, chased after the man.
(E) Five neighbors joined the pursuit when the man fled with the child's mobile phone.



Answer: B

Figure 49: An example of the task “Crime Abnormal”.

Comprehension - Scene - Disaster Understanding

Question:

Does the event's impact appear to have significantly disrupted the surrounding environment and infrastructure?

Options:

- (A) The environment shows very minor alterations with slight disruptions to traffic flow.
- (B) Slight disruptions in the environment are apparent, but infrastructure remains largely functional.
- (C) The event caused visible large-scale destruction to both the environment and infrastructure.
- (D) No observable impact on the environment or infrastructure.
- (E) Only the environment experienced significant disruption, but infrastructure is mostly intact.



Answer: C

Figure 50: An example of the task “Disaster Understanding”.

REFERENCES

- Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.
- Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.
- Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *European conference on computer vision*, pp. 609–623. Springer, 2012.
- Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *CoRR*, 2023.
- Tianyi Cheng, Dandan Shan, Ayda Sultan Hassen, Richard Ely Locke Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Jihoon Chung, Cheng-hsin Wu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13465–13474, 2021.
- Jieming Cui, Ziren Gong, Baoxiong Jia, Siyuan Huang, Zilong Zheng, Jianzhu Ma, and Yixin Zhu. Probio: A protocol-guided multimodal dataset for molecular biology lab. *Advances in Neural Information Processing Systems*, 36:41543–41571, 2023.
- Camille Dupont, Luis Tobias, and Bertrand Luvion. Crowd-11: A dataset for fine grained crowd behaviour analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 9–16, 2017.
- Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. Chalearn looking at people: A review of events and resources. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1594–1601. IEEE, 2017.
- Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.
- Shijia Feng, Michael Wray, Brian Sullivan, Youngkyoon Jang, Casimir Ludwig, Iain Gilchrist, and Walterio Mayol-Cuevas. Are you struggling? dataset and baselines for struggle determination in assembly videos, 2024. URL <https://arxiv.org/abs/2402.11057>.
- Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7214–7223, 2020.
- Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1343–1351, 2021.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.
- 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.
- Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.

- Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22072–22086, 2024.
- Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multiview dataset for learning multi-agent multi-view activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid–hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4572–4581, 2024.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL <https://aclanthology.org/D14-1086/>.
- Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1667–1675, 2017.
- Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE international conference on computer vision*, pp. 2650–2659, 2017.
- Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: A knowledge engine foundation for human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8494–8506, 2022.
- 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.
- 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*, pp. 740–755. Springer, 2014.
- Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.
- Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3566–3574, 2019.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Yu Luo, Jianbo Ye, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International journal of computer vision*, 128:1–25, 2020.
- Hu Ming, Wang Lin, Yan Siyuan, Ma Don, Ren Qingli, Xia Peng, Feng Wei, Duan Peibo, Ju Lie, and Ge Zongyuan. Nurvid: A large expert-level video database for nursing procedure activity understanding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, et al. Captaincook4d: A dataset for understanding errors in procedural activities. *Advances in Neural Information Processing Systems*, 37:135626–135679, 2024.
- Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Contiente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew

- Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=HYEGXFnPoq>.
- Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution.
- Mohammadreza Salehi, Jae Sung Park, Tanush Yadav, Aditya Kusupati, Ranjay Krishna, Yejin Choi, Hannaneh Hajishirzi, and Ali Farhadi. Actionatlas: A videoqa benchmark for domain-specialized action recognition, 2024. URL <https://arxiv.org/abs/2410.05774>.
- Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.
- Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8001–8013, October 2023.
- Hao Wang, Junchao Liao, Tianheng Cheng, Zewen Gao, Hao Liu, Bo Ren, Xiang Bai, and Wenyu Liu. Knowledge mining with scene text for fine-grained recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4614–4623. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00458. URL <https://doi.org/10.1109/CVPR52688.2022.00458>.
- Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 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 (ICCV)*, pp. 20270–20281, October 2023.
- Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzra, Zhuo Deng, Bilan Liu, James Z Wang, and Cheng-hao Kuo. Mebow: Monocular estimation of body orientation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3451–3461, 2020.
- Takuma Yagi, Misaki Ohashi, Yifei Huang, Ryosuke Furuta, Shungo Adachi, Toutai Mitsuyama, and Yoichi Sato. Finebio: a fine-grained video dataset of biological experiments with hierarchical annotation. *arXiv preprint arXiv:2402.00293*, 2024.
- Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23056–23065, 2023.
- Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018.
- Hao Zheng, Regina Lee, and Yuqian Lu. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. *Advances in Neural Information Processing Systems*, 36: 67069–67081, 2023.