Perception Test: A Diagnostic Benchmark for Multimodal Video Models

Viorica Pătrăucean[1*] DeepMind
Lucas Smaira DeepMind
Ankush Gupta DeepMind
Adrià Recasens Continente DeepMind
Larisa Markeeva DeepMind
Dylan Banarse DeepMind
Skanda Koppula DeepMind
Joseph Heyward DeepMind
Mateusz Malinowski DeepMind
Yi Yang DeepMind
Carl Doersch DeepMind
Tatiana Matejovicova DeepMind
Yury Sulsky DeepMind
Antoine Miech DeepMind
Alex Frechette DeepMind
Hanna Klimczak DeepMind
Raphael Koster DeepMind
Junlin Zhang DeepMind
Stephanie Winkler DeepMind
Yusuf Aytar DeepMind
Simon Osindero DeepMind
Dima Damen University of Bristol
Andrew Zisserman University of Oxford, DeepMind
João Carreira[1] DeepMind

Abstract

We propose a novel multimodal video benchmark – the *Perception Test* – to evaluate the perception and reasoning skills of pre-trained multimodal models (e.g. Flamingo, BEiT-3, or GPT-4). Compared to existing benchmarks that focus on *computational tasks* (e.g. classification, detection or tracking), the *Perception Test* focuses on *skills* (Memory, Abstraction, Physics, Semantics) and *types of reasoning* (descriptive, explanatory, predictive, counterfactual) across video, audio, and text modalities, to provide a comprehensive and efficient evaluation tool. The benchmark probes pre-trained models for their *transfer* capabilities, in a zero-shot / few-shot or limited finetuning regime. For these purposes, the *Perception Test* introduces 11.6k real-world videos, 23s average length, designed to show perceptually interesting situations, filmed by around 100 participants worldwide. The videos are densely annotated with six types of labels (multiple-choice and grounded video question-answers, object and point tracks, temporal action and sound segments), enabling both language and non-language evaluations. The fine-tuning and validation splits of the benchmark are publicly available (CC-BY license), in addition to a challenge server with a held-out test split. Human baseline results compared to state-of-the-art video QA models show a significant gap in performance (91.4% vs 43.6%), suggesting that there is significant room for improvement in multimodal video understanding.

Dataset, baselines code, and challenge server are available at `https://github.com/deepmind/perception_test`

[*]Corresponding author `viorica@google.com`, [1]shared senior contribution

1 Introduction

Significant progress in multimodal models has been made recently due to large-scale training on multimodal data. Models like Flamingo [4], PerceiverIO [33], BEiT-3 [48], GPT-4 [42] show remarkable versatility, dealing with diverse data sources and tackling new tasks by observing only a handful of examples. This is a major departure from specialised models that are typical in computer vision, *e.g.* image or action classifiers [51, 20], object detectors [14], or object trackers [46], opening up the path towards general perception and reasoning models.

Benchmarking these models in a robust and efficient way is key in expanding their capabilities, by allowing researchers to rank model design and training choices and identify areas for improvement. Many perception-related benchmarks exist, for example Imagenet for image classification [17], Kinetics for video action recognition [36], Audioset for audio event classification [25], TAO for object tracking [16], or VQA for image question-answering [28], to name only a few. While these benchmarks have led to amazing progress, they all target restricted aspects of perception, focusing on specific computational tasks: *e.g.* image benchmarks discard the temporal dimension, visual question-answering tends to focus on only high-level semantic scene understanding, and object tracking focuses on lower-level, texture-based cues. Gluing several datasets together to benchmark more general models (as is done in Flamingo, PerceiverIO, BEiT-3, or GPT-4) improves coverage, but results in an expensive evaluation procedure that still misses important general perception abilities, *e.g.* physics understanding or memory. Few existing benchmarks even define tasks over both audio and visual modalities [29], much less more complex combinations of modalities and tasks. Furthermore, most prior work provides large training sets and thus benchmark models for in-dataset capabilities.

In this work, we propose the *Perception Test* – a benchmark formed of purposefully designed, filmed, and annotated real-world videos that aims to comprehensively assess the capabilities of multimodal perception models across different skill areas (Memory, Abstraction, Physics, Semantics), types of reasoning [52] (*descriptive*, *explanatory*, *predictive*, and *counterfactual*), and modalities (video, audio, text). Our benchmark draws inspiration from diagnostic synthetic datasets like CATER [26] or CLEVRER [52], behavioral tests like the Visual Turing Test [40, 24], experiments in developmental psychology [1, 6, 31], and motor-free perception screening tests used for children or adults [41, 22].

To avoid benchmark overfitting, we propose a generalisation-focused evaluation regime. We aim to benchmark any representation or model, pre-trained with any *external* dataset or task, of any scale available. The *Perception Test* itself contains a small training set, intended for fine-tuning task decoders or prompting the model, and the rest is used for evaluation (public validation and held out test sets). In this regime, we can more robustly assess the *transfer* abilities of these models, such that improvement on the benchmark can more reliably predict generalisation to real-world operation.

The dataset contains 11.6K real-world videos, densely annotated with 190K object and 8.6K point tracks, 73.5K temporal action segments, 137K temporal sound segments, 38K multiple-choice video question-answer (mc-vQA) pairs and 6K grounded video question-answer (g-vQA) pairs, enabling both language and non-language evaluations, to ensure a thorough assessment; see Figure 1 and Table 3. Having multiple types of annotations per video is useful also for analysis purposes, as the correlations between successes and failures across tasks may uncover biases that prevent generalisation. For example, if a model correctly classifies an action, but then cannot localise where in space that action occurred, this may point to an incorrect understanding of the scene. Furthermore, fewer videos with a higher density of annotations enables efficient evaluation, as latent representations may be shared across tasks.

We open-source the videos and annotations in the training and validation splits. An evaluation server is made available together with the videos from the held-out test split. Since currently there is no model that can tackle all the evaluation tasks in our benchmark, we provide baseline results for per-task models: object tracking, point tracking, temporal action localisation, temporal sound localisation, multiple-choice video question-answering, and grounded video question-answering. For the mc-vQA task, the performance is mapped across skill areas (memory, abstraction, physics, semantics), and types of reasoning (descriptive, explanatory, predictive, counterfactual).

In the next section (section 2), we discuss related work in more detail, highlighting what sets the *Perception Test* apart in the rich landscape of multimodal benchmarks. In sections 3 and 4, we describe the videos and annotations in the *Perception Test*, with details about the diversity of participants involved in filming the videos. In section 5, we introduce the computational tasks enabled

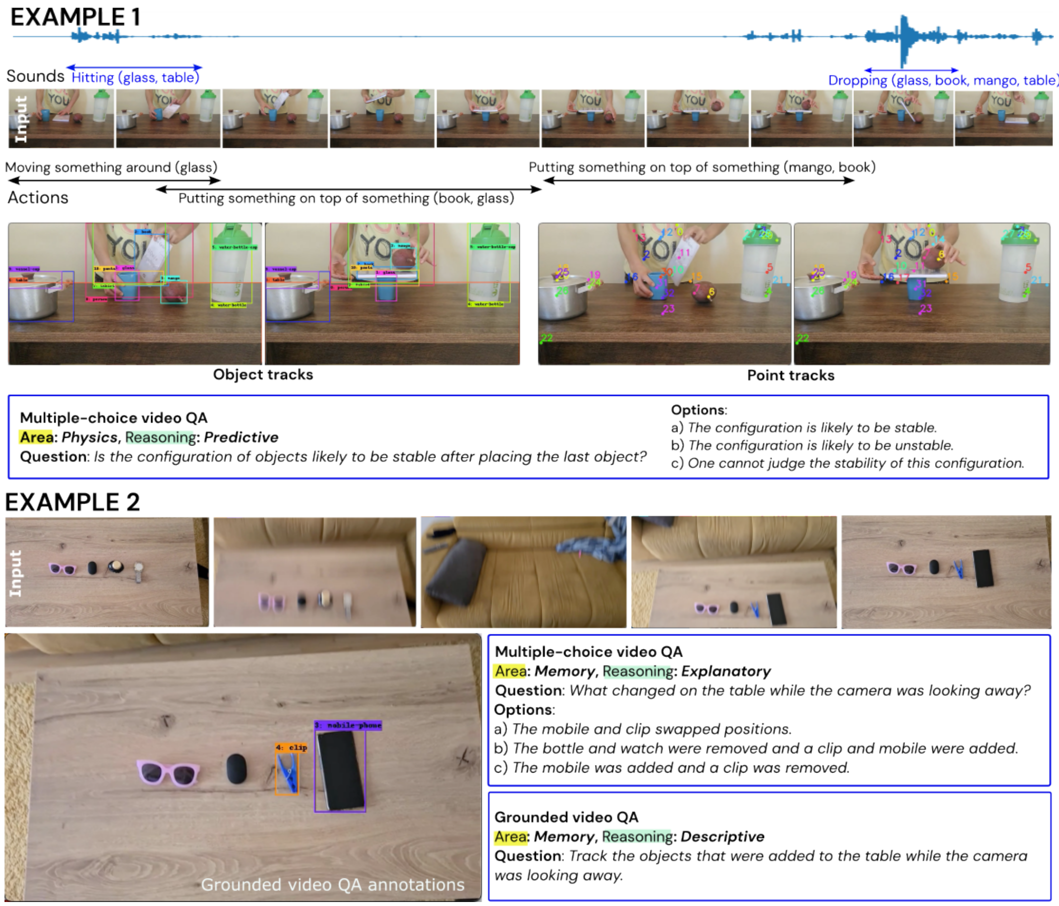

Figure 1: The *Perception Test* at a glance: 6 types of annotations (object & point tracks, action & sound segments, multiple-choice videoQA and grounded videoQA) and tasks spanning different skill areas and types of reasoning; see more examples in the supplementary material.

by these annotations, together with evaluation metrics. In section 6, we discuss per-task results obtained using baselines from the literature. We also include preliminary results from a human baseline. We conclude with a summary and directions of future work in section 7.

2 Related work

A large number of perception-related benchmarks exist in the literature, covering various computational tasks or modalities. We focus the discussion here on video benchmarks and highlight the differences between the *Perception Test* and prior work, in terms of data collection process, covered modalities, and available annotations and tasks.

Existing real-world benchmarks rely on one of the following data sources: **(i)** Videos collected from the web or repositories like Youtube, *e.g.* Kinetics [36], ActivityNet [10], VGGSound [12], HVU [18], ActivityNet-QA [53], tGIFQA [34]; **(ii)** Videos collected on demand, filmed by volunteers doing arbitrary activities in indoor or outdoor scenes, *e.g.* EPIC-KITCHENS [15], Ego4D [29]; **(iii)** Videos collected on demand, filmed by crowd-sourced participants doing actions described in pre-defined scripts, mostly in indoor scenes, *e.g.* Charades [45], Something-Something v2 (SSv2) [27].

Invariably, all real-world benchmarks use crowd-sourced annotations to enable various computational tasks like action classification, object detection, or video captioning, to name only a few.

Annotating publicly available videos is useful for training. However, using this approach for general perception evaluation has multiple drawbacks. Large quantities of data would need to be amassed and carefully filtered and annotated to accumulate (statistically) sufficiently diverse samples showing

Dataset	Source	Skills	# videos	Dens	L(s)
Charades	C,R	S	10,000	14	30
SSv2	C,R	AS	108,499	1	4
Ego4D-v2	R	MS	205,534‡	9*	492†
CLEVRER♭	C,Y	P	60,000	N/A	5
Perception Test	C,R	MAPS	11,620	761	23

Table 1: Characteristics of different datasets compared to the *Perception Test*. Dataset sources: Scripted (C), Real (R) and Synthetic (Y). Skill areas: Memory (M), Abstraction (A), Physics (P), Semantics (S). Dens: Average number of annotations per video. L: Average video length in seconds. ‡number of annotated clips, *reported for hand-objects subset with the highest density of annotations, †reported for ELM NLQ subset with highest average clip length. ♭: Annotations are extracted directly from the simulator.

perceptually interesting situations that require skills like memory, abstraction, physics and semantics understanding. In addition, some types of data are simply not available, *e.g.* situations showing incorrect execution of simple tasks like tying shoe laces. As we aim to assess more diverse skills, we chose to design video scripts that show perceptually interesting and diverse situations and film these with crowd-sourced participants from different places in the world to ensure diversity of video content and appearance. Different from Charades where the scripts were designed by crowd-sourced workers, our scripts are designed by our research team, similar to Something-something (v2). However, we did not aim to obtain an exhaustive coverage of simple actions like in SSv2. Instead, we designed more complex scripts, containing multiple actions, to probe for more advanced reasoning skills beyond action classification.

A few research works have highlighted the need for robust diagnostics benchmarks, *e.g.* CATER [26], CLEVRER [52], IntPhys [44], Physion [7]. Their authors developed synthetic datasets to evaluate in a more systematic way, across different levels of difficulty, models' abilities to reason about intuitive physics (object collisions, motion, object permanence). We share the same motivation of creating a diagnostic test, and we aim to cover aspects related to memory, abstraction, intuitive physics, and semantics, using real-world videos. To achieve this, in addition to designing the video scripts, our research team also designed the questions for each script type for the high-level tasks (mc-vQA and g-vQA); the answers per video were provided by crowd-sourced annotators.

Table 1 summarises the characteristics of the *Perception Test* compared to previous efforts. It can be observed that the *Perception Test* has a better coverage of skill areas and higher density of annotations[2]. Size-wise, the *Perception Test* is comparable to Charades, but much smaller than Ego4D or SSv2. We emphasise that the *Perception Test* is not designed to be a large-scale training dataset. Instead, it is an evaluation benchmark, with limited fine-tuning data, meant to assess the transfer capabilities of models.

3 Videos in the *Perception Test*

Inspired by how human perception screening tests are carefully designed by experts in developmental psychology or medicine (*e.g.* [13]), we designed video scripts and tasks to diagnose the perception skills of our models.

Scripts design: Our goal was not to obtain an exhaustive coverage of activities or types of scenes. Instead, we selected four areas – Memory, Abstraction, Physics, Semantics – within which several skills should be tested (see Table 2, second column) through tasks that require different types of reasoning: descriptive, explanatory, predictive, or counterfactual [52]. The skills selection took into account blind spots of existing benchmarks, weaknesses of current models, and aspects that are important for real-world scene understanding.

We then created scripts describing simple situations or games that can be easily performed by any one person (non-professional actor) using the items available in a regular household, or items that can be easily crafted if not available (*e.g.* letters or geometric shapes crafted from paper or cardboard). Each script consists of a brief description of the scene, followed by a description of the actions to be

[2] We count every labeled box, point, temporal segment, or question as a separate annotation

performed, together with specification of the camera placement (static camera one viewpoint; static camera 2 viewpoints; static camera and moving camera). To enhance content diversity, each script had considerable room for variability in the number of objects to be included in the scene or types of actions to be performed, or order of actions.

We prioritised situations where we can test high-level concepts like memory through low-level tasks like object tracking and the other way around: low-level physics understanding probed through high-level tasks like question-answering. In addition, we included in each script elements that could make the situations more interesting and challenging. For example, in cooking scripts (*e.g.* making tea, making salad), we added *distractor actions* [45], i.e. actions not relevant for making tea and that have no impact on the outcome of the making tea sequence, like clapping hands, or hitting a kettle with a spoon; this allows probing for understanding of causal relations between actions. We also included *distractor objects* in the scene description, i.e. objects that are not relevant for the current script, but which are relevant for other scripts, like tomatoes present on the table during the make tea activity [47]. For all the scripts, we also asked participants to include in the scene some *adversarial configurations of objects e.g.* a shoe on the table. This allows us to probe models for understanding of spatial relations of objects when the language biases are not valid. Finally, some of the script variations include *adversarial actions* [27], i.e. incorrectly executed actions. For example, when making the tea, all the steps are done normally, but one is incorrectly executed, like pouring water from an empty kettle. In this way, we can probe for understanding of task completion, in a more complex setup than the adversarial action classification used in SSv2 dataset [27].

Table 2 and Figure 1 show examples of situations included in the scripts to probe for different skills in the four areas and different types of reasoning. Note that the videos associated with a script allows defining tasks and questions across multiple skill areas. All-in-all, we designed 37 scripts, each with 2-5 variations, to obtain a diverse dataset. Having multiple variations per script allows us to ask the exact same question with the same set of options, and the correct answer depends on the specific script variation – in this way, we can avoid language biases in questions that give away the answer [37].

Video filming: Ensuring diversity of participants and scenes depicted in the videos was a critical consideration when developing the benchmark. Using a crowdsourcing pool, we selected around 100 participants from different countries of different ethnicity and gender and aimed to have a diverse representation within each video script. We include in the supplementary material details about the self-reported demographics of participants. Each script variation was filmed by at least a dozen of different participants, using most often a mobile-phone camera, resulting in high-resolution audio-visual assets. For scripts to be filmed from two different viewpoints, the recording was most often done sequentially by repeating the script; a few participants recorded simultaneously using two filming devices. About 15% of the videos were filmed with a moving camera. Most of the videos were filmed indoors in the living room or kitchen, with a small number being filmed in the bathroom or outdoors (about 1%). Most of the activities are performed on a tabletop, but some are also performed on the floor or on a chair. To avoid privacy concerns, we instructed the participants to not record their faces or voices. This does not constitute a limitation of the dataset since the focus in our scripts is on object interactions. The participants gave their consent for the data to be used, published, and stored for perpetuity.

Splits: The *Perception Test* contains 11609 videos (with audio), 23s average length. It is divided into a small training split (2184 videos, $\sim 20\%$ of the data) that can be used for fine-tuning or prompting, a validation split (5900 videos, $\sim 50\%$ of the data), and a held-out test split (3525 videos, $\sim 30\%$ of the data) available through the evaluation server.

We optimised to obtain a good balance across all annotation types and camera motions across the 3 splits; see section 6 in the appendix.

4 Annotations in the *Perception Test*

We annotate these videos with multiple types of annotations to cover low-level and high-level aspects, spatial and temporal, and enable language and non-language evaluations: object and point tracks, temporal action and sound segments, multiple-choice and grounded video question-answers. We include a summary of the number of annotations of different types in Table 3 and visualisations in Figure 1.

Skill area	Skill	Example of situations and questions or tasks
Memory	Visual discrimination	Objects are shown in front of the camera, with some shown more than once. **Task**: Detect which objects were shown multiple times.
	Change detection	The camera is filming a table, then looks away for a few seconds, then looks back at the table. Some changes may have occurred. **Task**: Explain what changed.
	Sequencing	Objects are put in a backpack. **Task**: List their order.
	Event recall	A person indicates a region on the table with the hand, then puts objects inside and outside the region. **Task**: List the objects put inside the region.
Abstraction	Object, action & event counting	A person turns a lamp on and off. **Task**: Count the number of times the illumination changed in the scene.
	Feature matching	A person puts wooden letters on the table. **Task**: List the letters that have the same colour.
	Pattern discovery	Geometric shapes are shown in a pattern. **Task**: Predict what shape would be shown next.
	Pattern breaking	A person puts multiple cups all facing upwards and one facing downwards. **Task**: Indicate the object that breaks the pattern.
Physics	Object permanence	A person plays a cups-game with 3-4 cups by hiding a small object under one of the cups, then shuffles the cups. **Task**: Predict where is the hidden object after shuffling.
	Spatial relations & containment	A person puts a bookmark in a book, then puts the same or another book in a backpack. **Task**: Where is the bookmark at the end?
	Object attributes	A person writes on a piece of paper. **Task**: Is the paper lined or plain?
	Motion & occluded interactions	A person moves an occluder object in front of a small object, sometimes moving also the small (occluded) object. **Task**: Was the small object moved?
	Solidity & collisions	A person launches objects against a blocker object, sometimes removing the blocker. **Task**: Does the object fall off the table?
	Conservation	A person pours an equal amount of water in 2 identical glasses, then pours all or part of the water from one glass in a taller or wider glass. **Task**: How much water is in the last glass?
	Stability	A person puts objects on top of each other in a stable or unstable configuration. **Task**: Predict if the configuration will be stable after placing the last object.
Semantics	Distractor actions & objects	A person makes tea, and does also some distractor actions unrelated to making tea, *e.g.* rotating a knife. **Task**: Identify the distractor action(s).
	Task completion & adversarial actions	A person ties shoe laces, but sometimes pretends to tie, or ties the lace of one shoe to the lace of the other shoe. **Task**: Detect if the action is done correctly.
	Object & part recognition	A person conceals a small object in one of their hands, then shuffles the hands. **Task**: Identify in which hand is the object held.
	Action & sound recognition	All scripts. **Task**: Detect the actions and sounds in the video from a pre-defined list.
	Place recognition	All scripts. **Task**: Detect where is the action taking place.
	State recognition	A person uses an electric device. **Task**: Indicate if the device is on.
	General knowledge & Language	Some objects are shown to the camera, some multiple times. **Task**: Given a list of arbitrary statements or word puzzles, some requiring general knowledge to solve, select the statement that contains a reference to the second distinct object shown.

Table 2: Examples of scripts probing for different skills in the four areas in the *Perception Test*.

Object tracks: Object tracks represent the *root annotation* of our benchmark. All the other annotations, except for multiple-choice vQA, are linked or grounded into object tracks. In the annotation process, we instructed annotators to focus on the objects that the person interacts with and the objects that are in the immediate vicinity of the area where the person is performing actions, which act as distractor objects. We annotated boxes at 1fps throughout the video, which gives a good trade-off between density of annotations and annotation cost. When the objects are occluded, the annotators marked an approximate position of the boxes. Some ambiguous classes still remain, like liquids being poured or objects being torn. The object names were defined from an open vocabulary. The

Annotation type	# classes	# annot	# videos	Rate (fps)
Objects tracks	5101	189940	11609	1
Point tracks	NA	8647	145	30
Action segments	63	73503	11353	30
Sound segments	16	137128	11433	30
mc-vQA	132	38060	10361	NA
g-vQA	34	6086	3063	1

Table 3: Annotations collected for the *Perception Test*. Each object or point track contains frame-level annotations at a certain *rate*, *e.g.* each point is annotated on every frame, at 30 fps. Action and sound segments are annotated at the original video frame rate. # classes refers to the number of unique object names for object tracks and the number of unique questions for mc-vQA and g-vQA.

annotators typically included object attributes as well (colour, material), resulting in a large number of unique names. A list of the most frequent words (object or attributes) in the benchmark is included in the supplementary material, Fig. A2 (left), together with the distribution of object tracks into various categories, *e.g.* objects involved in actions or sounds correlated with camera motion (Table A1).

Cups-game subset: We isolate the videos corresponding to the cups-game scripts, as they can be an interesting subset for probing object trackers' abilities to reason about motion, object permanence, or occluded interactions when different factors may influence the difficulty of the task, *e.g.* identical vs non-identical objects used in the game, transparent vs non-transparent objects, or number of objects used. This subset contains 598 videos, with 483 videos where the cups are identical, and 113 videos where the cups are transparent. Most of the videos have 3 cups (451 videos), 132 videos have 2 cups, and 34 videos have 4 cups. We also provide a visibility mask for each video showing when the hidden object is occluded.

Point tracks: Although object tracks based on bounding boxes allow probing some physical properties of objects, such as object permanence, solidity, and coarse motion, they don't fully describe articulated or non-rigid objects, thin objects that are not axis-aligned, or out-of-plane rotation. A better understanding of physical interactions arises if we can track how object *surfaces* move and deform over time as the interaction takes place. To this end, we annotate point tracks on object surfaces following the protocol of TAP-Vid [19]. Annotators were instructed to select points spanning all the different parts of the objects labelled in the object tracking task. Thus, each point is linked to one of the tracked objects. Points that are occluded are simply marked as occluded and not tracked. For translucent objects (*e.g.* glass cups), we only consider points to be 'visible' if they belong to the surface closest to the camera. The annotated points are sparse in space but dense in time. Table A2 included in the supplementary material gives the distribution of points that are moving or static, as well as those on videos with moving cameras.

Action segments with action-relevant objects: To capture temporal understanding and enable grounding over time, we annotate the videos with temporal segments belonging to a fixed set of templated labels, *e.g. putting something into something*, similar to [27]. These are associated with action-relevant object tracks, i.e. objects involved in the action. The action boundaries are defined based on contact with action-relevant objects. For instance, when a person puts sugar in a tea, the *putting something into something* action starts when the person picks up the spoon and ends when the person puts down the spoon. If, after putting the sugar, the person starts stirring with the same spoon, this defines a new segment as the type of action changed. The frequency of actions across the entire dataset is included in the supplementary material, Fig. A2 (right).

Sound segments with sound-relevant objects: Similarly to the action segment annotations but applied to the audio modality, we collect sound segment annotations grounded in object tracks. By watching the video and listening to the audio, the annotators define temporal sound segments and label them from a list of 16 audio segment labels. For each sound, the annotators also identify the object (or objects) involved in making the sound, or specify that these are out of the camera's field of view. For example, if an object is placed on the table making an audible sound, then both the object track and the table track are associated with the sound segment. The frequency of sounds across the entire dataset is included in the supplementary material.

Question-answers for video-level reasoning: Different from the existing VQA datasets, which rely on crowd-sourced questions and answers, we designed ourselves the questions per script to cover different types of reasoning [52]: descriptive, explanatory, predictive, counterfactual, and to cover

Area	# videoQA	Reasoning	# videoQA
Memory	7256 (36)	Descriptive	31536 (106)
Abstraction	12737 (58)	Explanatory	4513 (14)
Physics	23741 (80)	Predictive	1278 (7)
Semantics	24965 (82)	Counterfactual	733 (5)

Table 4: Number of videoQA pairs and (unique questions) per area and type of reasoning. Note that one question may be counted in multiple areas if it tests more than one skill. Each question is assigned a unique type of reasoning.

aspects that are important for operating in the real world, *e.g.* understanding task completion, detecting changes, and so on. The answers for all the questions per video were provided by crowd-sourced participants. As we are interested in non-ambiguous evaluation, we favour the multiple-choice setup over the open-language answer setup. To define challenging negative options, we partly relied on human annotators, partly sampling from the correct answers of other videos in the same type of script. Table 4 and Figure A4 (in the supplementary material) show the distribution of question-video pairs into perception skills, skill areas, and type of reasoning.

Question-answers with answer-relevant objects: As another way to connect high-level and low-level scene understanding capabilities, we define questions or tasks in language form, with answers given as object tracks. Similar to the regular question-answers above, these grounded questions are associated with skill areas and types of reasoning; most of these tasks fall under the Physics area.

5 Computational tasks in the *Perception Test*

We defined six computational tasks based on the annotations available in the *Perception Test*. We summarise in Table 5 the task definitions: input, output, and metric. It can be observed that the *Perception Test* combines lower-level dense prediction tasks like object and point tracking, whose outputs are box and point trajectories, with higher-level tasks like video question answering. For all the tasks, the video and audio are available as inputs, together with a task specification where applicable, *e.g.* the coordinates of a box to track for object tracking, or a language question and options for multiple-choice videoQA. Note that many other computational tasks can be defined based on the available annotations, *e.g.* video object detection, grounded temporal action/sound localisation, and more.

Single object tracking: In this task, the model should separately track every single object box labelled in the dataset starting from one of its first frames. In some cases ($\approx 20\%$) where the object is entering the field of view at the beginning or during the video, the first box may span only a few pixels, so it does not contain a representative view of the object. To deal with this problem, we use a heuristic to select a later frame, when the object is not touching the image boundary, to identify the query box for each object track. Performance is evaluated using the standard *average intersection-over-union* (IoU) metric, (also called average overlap), for evaluating long-term tracking without tracker re-initialization. It is defined as the average IoU over the entire track between the predicted and the ground-truth boxes [32, 11]. We also provide code for more fine-grained analysis, *e.g.* performance on objects in videos shot with static vs. moving cameras, objects involved in actions etc.

Cups-game subset: For the occluded object involved in cups-games, we use intersection as a metric for tracking (as opposed to Intersection-over-Union), to deal with the uncertainty of the position when the object is occluded.

Single point tracking: In this task, given a set of ground truth initial 2D point coordinates, the model should separately trace their spatial trajectories throughout the video. Performance is evaluated using the recently proposed *average Jaccard* metric for evaluating both long-term point tracking position and occlusion accuracy. This metric checks how similar the predicted and the ground-truth point tracks are, based on the average number of true positive matches, divided by the sum of true positives, false positives, and false negatives over the entire track [30, 19].

Temporal action / sound localisation: We define these two tasks similarly, as temporal segment detection problems. Given a video, the model predicts potentially overlapping temporal 1d-segment covering the actions/sounds and classifies them using a fixed set of labels. Performance is evaluated

Task	Output	Metric
Object tracking	box track	Avg. IoU
Point tracking	point track	Avg. Jaccard
Temporal action localisation	list of action segments	mAP
Temporal sound localisation	list of sound segments	mAP
mc-vQA	answer (1 out of 3)	top-1 accuracy
g-vQA	list of box tracks	HOTA

Table 5: Computational tasks in the *Perception Test*: the model receives a video with audio, plus a task-specific input (*e.g.* the coordinates of a bounding box for the object tracking task), and produces a task-specific prediction, evaluated using dedicated metrics.

Object Tracking	All	Static camera	Moving camera
all objects	0.66 / 0.67	0.70 / 0.69	0.42 / 0.54
action objects	0.48 / 0.53	0.50 / 0.54	0.31 / 0.47
sound objects	0.56 / 0.60	0.58 / 0.61	0.40 / 0.53
g-vQA boxes	0.38 / 0.50	0.43 / 0.51	0.26 / 0.47

Table 6: Static dummy baseline / SiamFC results, measured as average IoU, across different categories of objects in the *Perception Test*. Since many objects are static, the performance of the dummy baseline is good overall, but it degrades considerably when motion is involved, whereas the SiamFC tracker is more robust.

using the standard mean AP over classes [54] based on temporal IoU between predicted and ground truth temporal segments.

Multiple-choice video question-answering: In this task, the model receives, in parallel with the video, a question and three possible answers, out of which only one is correct, and the model has to pick one answer (33% random chance). For most of the questions, watching the video and reading the question are enough for providing a correct answer. A limited number of questions are formulated in a generic way, so the options are necessary for choosing the answer: *e.g. Which of the following statements describes the scene better?*. In some cases, choosing the answer by elimination of the false options may be simpler. Performance is evaluated by measuring top-1 accuracy. For a couple of scripts, the videos must be trimmed to not reveal the answer: in the cups-games and stable configurations videos, we provide a frame id where the video should be trimmed. For the train and validation splits we release the entire videos together with the cut frame id information. In the held-out test split, only the trimmed videos are available for these particular video types.

Grounded video question-answering: This task is similar to conditional multiple-object tracking, with the conditioning given as a language task or question as opposed to a class label [38]. The answers are object tracks defined throughout the video and we use HOTA [39] metrics to evaluate performance. In some situations, the initial parts of the track might not be relevant for the question, *e.g. Track the object that was removed from the table* and the object is removed halfway through the video. However, given that we do not enforce causal processing of the video, the track prediction for the initial part can still be done in hindsight.

6 Baselines

Ideally, a single model should be able to perform all the tasks in the *Perception Test*, in a zero/few-shot setting or by fine-tuning on our limited training set. Since such a model is not available in the literature, we include results obtained with per-task baselines on the validation split for all the six tasks in the *Perception Test*.

For the multiple-choice video QA, we also include the results of a human baseline.

Object tracking: We report baseline results using the SiamFC model [8] (UniTrack [49] implementation). SiamFC was chosen due to its high-performance on a number of single-object tracking benchmarks when running in zero-shot setting [21]. The results for the different categories of objects (involved in actions or in sounds, etc) are included in Table 6, aggregated based on camera motion.

Point tracking: We report baseline results using a TAP-Net model [19] trained on Kubric [30] and transferred zero-shot. The model operates on 256x256 resolution (aspect ratio is not preserved) and consumes the whole video directly. Table 7 shows the results. As expected, both moving points and points seen through a moving camera are considerably harder to track.

Point tracking	All points	Static camera	Moving camera
static baseline	0.361	0.410	0.088
TapNet	0.401	0.414	0.328

Table 7: Static dummy baseline vs TapNet [19] results (average Jaccard, higher is better) for the point tracking task on the validation set. Note that the TapNet model was not trained on this benchmark, it was evaluated zero-shot. Please see more details in supplementary material.

Temporal action localisation: We obtained baseline results for temporal action localisation using ActionFormer [54] with different pretrained features: TSP video features from [5] pre-trained on ActivityNet, MMV audio features from [2] pre-trained on AudioSet, and a multimodal input by concatenating the video and audio features. We trained the transformer blocks and the classification and regression heads to accommodate for the number of classes included in our dataset. The resulting mean average precision is included in Table 8, top. The baseline struggles mostly with rare action classes and pretend actions, which are confused with their non-pretend counterpart class. Using only the audio modality leads to very poor performance, whereas using multimodal inputs does not increase the performance significantly. More details are given in the appendix.

		Temporal Action Localisation						
Model	**Modality**	**@0.1**	**@0.2**	**@0.3**	**@0.4**	**@0.5**	**Avg**	**# epochs**
ActionFormer	video	17.67	16.56	15.13	13.28	11.07	14.74	35
ActionFormer	audio	7.25	6.53	5.70	4.67	3.64	5.56	55
ActionFormer	video+audio	18.82	17.63	15.98	13.99	11.37	15.56	35

		Temporal Sound Localisation						
Model	**Modality**	**@0.1**	**@0.2**	**@0.3**	**@0.4**	**@0.5**	**Avg**	**# epochs**
ActionFormer	video	17.85	15.54	13.81	12.11	5.89	13.04	55
ActionFormer	audio	16.28	13.58	10.80	8.43	5.87	10.99	80
ActionFormer	video+audio	22.24	18.99	15.36	11.99	8.74	15.46	55

Table 8: Mean average precision (mAP) for temporal action localisation (top) and sound localisation (bottom) tasks using ActionFormer as baseline. IoU for 0.1-0.5 are averaged as in [15]. # epochs represents the number of training epochs used to obtain the best results for each experiment setup.

Temporal sound localisation: We use the same model architecture and pre-trained features as above. We trained from scratch the transformer blocks and the classification and regression heads. For both training and evaluation, we keep only 11 sound classes, excluding the classes corresponding to indistinguishable sounds (*e.g. Other:background*, *Other:human*), as they hinder learning. The resulting mean average precision is included in Table 8, bottom. The best performance is obtained when features from both modalities are used as input. More details are included in the appendix.

Multiple-choice videoQA: For this task, we report the results of Flamingo [3] – a state-of-the-art video-language model, together with a dummy frequency-based baseline and a human baseline; see Table 9, Figure 2 and 3.

Frequency baseline. Given that we have a fixed set of question-answer pairs defined over multiple videos, we compute a simple baseline that always picks the most frequent correct answer (on the training set) during evaluation; this baseline obtains 47%. One can also compute this baseline on a random subset of training examples for each question type, see Table 9. This is a fairer dummy baseline for models using few-shot evaluation.

Human baseline. We ran a small study for the mc-vQA task with human participants. We used 126 questions from the dataset, with one video per question selected at random. We recruited 30 crowd-sourced participants (half male, half female, with advanced English skills), different from the raters annotating the videos. Each participant answered a subset of 42 questions, resulting in 10 answers per question. The performance per area and type of reasoning is detailed in Figure 2. The overall average accuracy was 91.4%. The mistakes occurred in situations difficult to judge from the given viewpoint, *e.g.* if a configuration of objects would be stable (without seeing the end of the video), or in edge cases where humans overlooked details happening very early on in the video. It is worth noting that running the study was straightforward, the participants did not require any training, similar to a zero-shot setup. The median time spent to answer 42 questions was 30 minutes.

Flamingo. We run the model with a maximum of 30 frames sampled at 1fps, spatial resolution 320. When the videos are longer than 30 seconds, we use only the middle clip. The audio modality is ignored as the original model was not trained to deal with it. The different options are scored based on likelihood. We considered zero-shot and 8-shot settings. In the zero-shot setting, the smaller version of the model obtains 43.6% on the test set. In the 8-shot setting, we sample 8 examples and associated ground truth responses from each question in the training set and use as prompts. The resulting accuracy is 45.8%, again obtained by the smaller version of the model. Flamingo struggles the most on counterfactual questions, where its accuracy is below random, pointing to an important ability that our models clearly lack [43]. Interestingly, the larger versions of the model (due to larger language branches) seem to fare worse, which suggests that the difficulty of the tasks is on the vision side, not the language side. Figure 3 details the performance across skills compared to random baseline. It can be observed that on several skills, e.g. (Piaget) conservation task or change detection, Flamingo is below random.

mc-vQA	0-shot	8-shot	Full
Flamingo-3B	43.6	**45.8**	-
Flamingo-9B	40.5	44.4	-
Flamingo-80B	41.6	45.4	-
Frequency	33.3	43.0	**47.0**
Human	**91.4**	-	-

Table 9: mc-vQA top-1 accuracy (higher is better), for different evaluation modes and different models, including a human baseline, on the validation split. The frequency baseline picks the most frequent answer in the training set (for zero-shot this corresponds to random chance). "-" refers to numbers that were not collected.

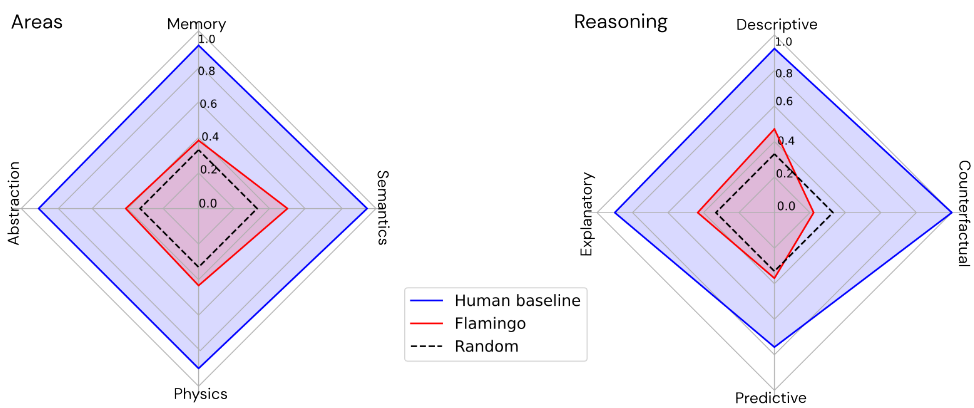

Figure 2: Zero-shot human baseline results for a subset of questions and videos in the *Perception Test*, compared to 8-shot Flamingo-3B results on the entire validation set. The black dashed line indicates the random baseline. Even in the 8-shot regime, Flamingo is far from the zero-shot human baseline and seems to struggle the most with memory-related skills and counterfactual reasoning.

Grounded video question-answering: In absence of a dedicated baseline in the literature for the type of grounded videoQA that we propose (input: text query, output/answer: object tracks), we obtain a simple baseline by running MDETR [35] on the middle frame of each video using the query as input, and then we use Stark tracker [50] to propagate the MDETR detections forward and backward in the video. We measure the performance of this baseline using HOTA metrics, which integrate detection, association, and localisation scores. As expected, the performance of this baseline is poor; see Table 10 and Figure A6 in the appendix. The failures are caused mainly by poor detection results – since the tasks are temporal in nature, extracting *seed* boxes from the middle frame is not

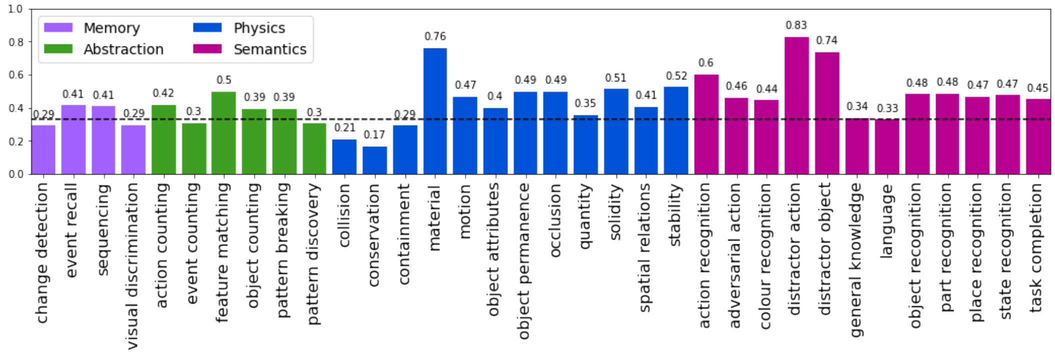

Figure 3: 8-shot Flamingo-3B performance on the validation set across skills. The black dashed line indicates the random baseline.

Model	HOTA	LocA	DetA	AssA
MDETR+Stark	0.1	0.68	0.03	0.33

Table 10: HOTA results on the validation split for the grounded vQA task in the *Perception Test*.

sufficient to solve the tasks, calling for models capable of dealing with both spatial and temporal dimensions.

7 Conclusion

We propose a diagnostic benchmark for multimodal models, that probes for memory, abstraction, physics, and semantic capabilities, across visual, audio, and text modalities, using real-world videos purposefully designed and filmed to show interesting perceptual situations. Solving the tasks requires different types of reasoning: descriptive, explanatory, predictive, and counterfactual. The videos are densely labeled with six types of annotations (objects and point tracks, action and sound segments, multiple-choice and grounded video question-answer pairs), which enables evaluating models across many different dimensions on a common set of videos. It also sets up a foundation for more advanced future tasks that combine these annotations in various ways.

We are open-sourcing the videos and the annotations in the train and validation splits, together with per-task baseline results and evaluation code. A challenge server is available, to evaluate models on the held-out test split. In principle, any model can be evaluated on our benchmark, either in a zero/few-shot setting or by fine-tuning on our limited train split. An ideal perception model would be able to perform all the tasks in our benchmark. We address ethical and societal aspects that our work may impact in the supplementary material. Our results suggest that state-of-the-art zero-shot video-language models do only slightly above a dummy frequency based baseline, whereas humans in the same setting are nearly perfect. This gives a new perspective on understanding models' limitations and could help narrowing down areas of improvement to guide research. In addition, by combining low-level and high-level annotations across multiple modalities and tasks, we hope to enable cross-pollination between communities that are currently fairly fragmented, *e.g.* the community working on tasks like tracking and flow estimation using benchmarks like KITTI [23] or Sintel [9], with the community working on high-level scene understanding using benchmarks like ImageNet, Kinetics, or VQA. Finally, we hope to collaborate with the community to continuously grow and improve this benchmark, by adding new videos, tasks, modalities, tool use, or even new languages, to build a comprehensive diagnostic test for multimodal perception models.

Acknowledgments

We are grateful to Luis Piloto, Kenneth Marino, Luyu Wang, Felix Hill, Martin Chadwick, Lucy Campbell-Gillingham, Boxi Wu, Drew Jaegle, Pauline Luc, Marianne Monteiro, Anna Bulanova, Radu Isac, Muqthar Mohammad, Vijay Vibha Tumala, Mahesh Maddinala, Yiwen Luo, Alina

Kuznetsova, Aida Nematzadeh, Lisa Anne Hendricks, Aishwarya Agrawal, Nando de Freitas, Matt Botvinick, Shane Legg, and Relja Arandjelovic for providing insightful input on this project.

References

[1] A. Aguiar and R. Baillargeon. Developments in young infants' reasoning about occluded objects. *Cognitive Psychology*, 45:267–336, 2002.

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=EbMuimAbPbs`.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL `https://arxiv.org/abs/2204.14198`.

[5] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. TSP: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 3173–3183, 2021.

[6] Renée Baillargeon. Physical reasoning in young infants: Seeking explanations for impossible events. *British Journal of Development Psychology*, 12:9–33, 1994.

[7] Daniel Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Fei-Fei Li, Nancy Kanwisher, Josh Tenenbaum, Dan Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d09bf41544a3365a46c9077ebb5e35c3-Abstract-round1.html`.

[8] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. *arXiv preprint arXiv:1606.09549*, 2016.

[9] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, 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 (CVPR)*, June 2015.

[11] Luka Čehovin, Aleš Leonardis, and Matej Kristan. Visual object tracking performance measures revisited. *IEEE Transactions on Image Processing*, 25(3):1261–1274, 2016.

[12] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[13] Deirdre M Cooke, Kryss McKenna, Jennifer Fleming, and Ross Darnell. The reliability of the occupational therapy adult perceptual screening test (ot-apst). *British Journal of Occupational Therapy*, 68(11):509–517, 2005.

[14] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7373–7382, June 2021.

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022.

[16] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. *Lecture Notes in Computer Science*, pages 436–454, 2020. ISSN 1611-3349. doi: 10.1007/978-3-030-58558-7_26. URL `http://dx.doi.org/10.1007/978-3-030-58558-7_26`.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020.

[19] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens Continente, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-vid: A benchmark for tracking any point in a video. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=Zmosb2KfzYd`.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[21] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Mingzhen Huang, Juehuan Liu, Yong Xu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129(2):439–461, 2021.

[22] Marianne Frostig and David Horne. *The Frostig program for the development of visual perception: Teacher's guide*. Follett Publishing Company in collaboration with Curriculum Materials ..., 1965.

[23] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[24] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual Turing test for computer vision systems. *Proceedings of the National Academy of Science*, 112(12):3618–3623, March 2015. doi: 10.1073/pnas.1422953112.

[25] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

[26] Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*, 2020.

[27] Raghav Goyal, Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, Ingo Bax, and Roland Memisevic. Evaluating visual "common sense" using fine-grained classification and captioning tasks, 2018. URL `https://openreview.net/forum?id=rkX9Z_kwf`.

[28] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Yu Heng Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar,

Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3761, 2022.

[31] Eileen Mavis Hetherington, Ross D. Parke, and Virginia Otis Locke. *Child psychology: A contemporary viewpoint, 5th ed.* McGraw-Hill, 1999.

[32] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1562–1577, 2019.

[33] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=fILj7WpI-g`.

[34] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video Question Answering with Spatio-Temporal Reasoning. *IJCV*, 2019.

[35] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021.

[36] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[37] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2776–2785, June 2021.

[38] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, and Arnold W. M. Smeulders. Tracking by natural language specification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7350–7358, 2017. doi: 10.1109/CVPR.2017.777.

[39] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, pages 1–31, 2020.

[40] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. *arXiv preprint arXiv:1410.8027*, 2014.

[41] Nancy A Martin and Morrison F Gardner. *Test of visual perceptual skills*. Academic Therapy Publications Novato, CA, 2006.

[42] OpenAI. Gpt-4 technical report, 2023.

[43] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

[44] Ronan Riochet, Mario Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 03 2018. doi: 10.1109/TPAMI.2021.3083839.

[45] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, Ivan Laptev, and Abhinav Kumar Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv*, abs/1604.01753, 2016.

[46] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv: 2012.15460*, 2020.

[47] Andrea Tacchetti, Leyla Isik, and Tomaso Poggio. Invariant Action Recognition Dataset, 2019. URL https://doi.org/10.7910/DVN/DMT0PG.

[48] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. URL https://arxiv.org/abs/2208.10442.

[49] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34:726–738, 2021.

[50] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10428–10437, 2021.

[51] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview Transformers for Video Recognition. *arXiv e-prints*, art. arXiv:2201.04288, January 2022.

[52] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

[53] 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 *AAAI*, pages 9127–9134, 2019.

[54] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 2022.

Appendix

1 *Perception Test* at a glance

Figure A1 and the presentation video available at `https://github.com/deepmind/perception_test` summarise the types of videos, annotations, and tasks available in the *Perception Test*.

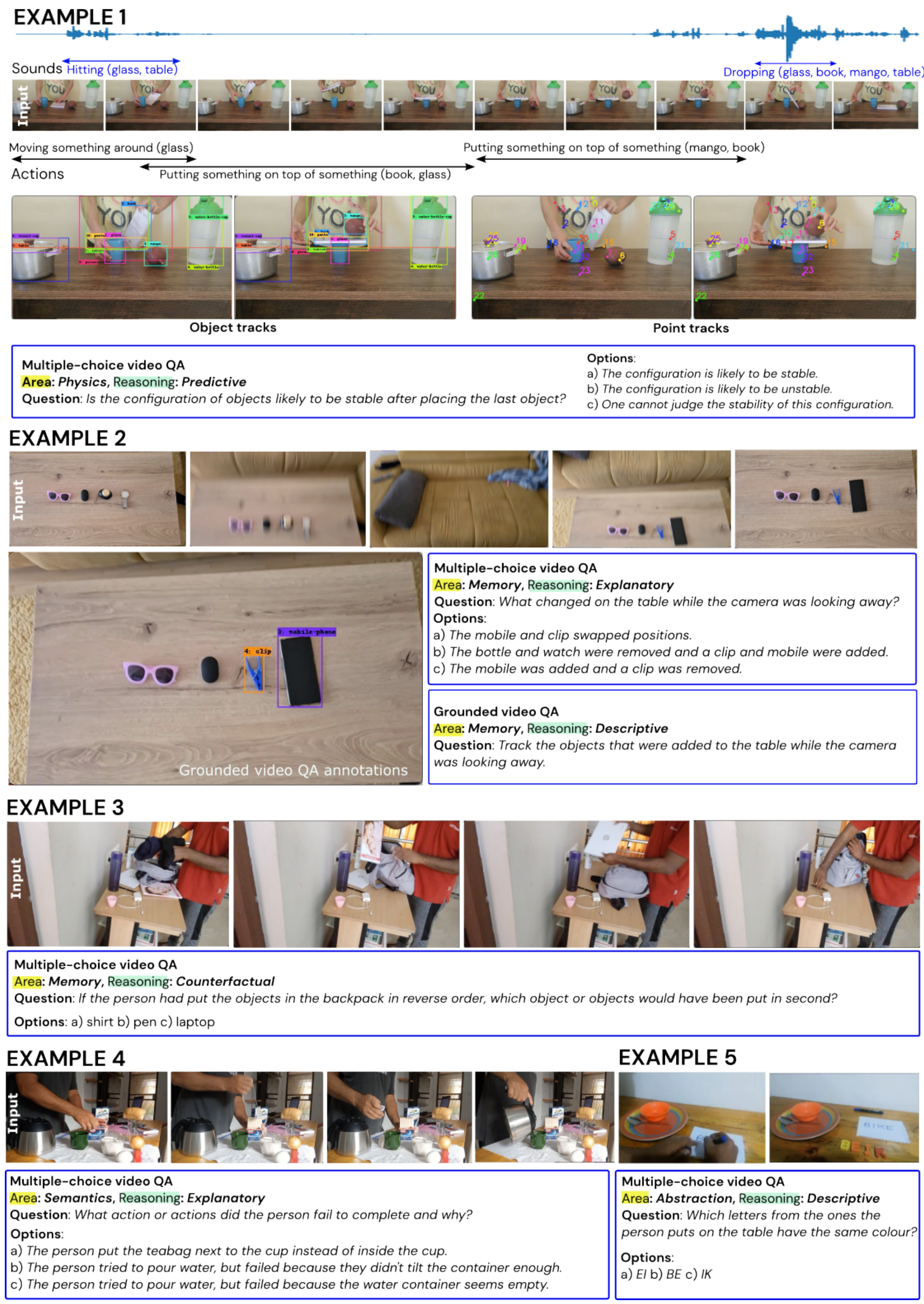

Figure A1: The *Perception Test* contains 6 types of annotations (object & point tracks, action & sound segments, multiple-choice videoQA and grounded videoQA) and tasks spanning 4 skill areas (Memory, Asbtraction, Physics, Semantics, and 4 types of reasoning (Descriptive, Explanatory, Predictive, Counterfactual). See the presentation video at `https://github.com/deepmind/perception_test` for more examples.

2 More details about annotations in the *Perception Test*

The distributions of object and point tracks across camera motion and objects involved in actions, sounds, and grounded vQA are included in Table A1 and Table A2. Figures A2 and A3 present the frequency of popular words included in object names, and the distribution of actions and sounds respectively. Figure A4 shows the distribution of questions across skills.

Camera	**Static**	**Moving**	**Total**
# total objects	165552	26164	191716
# action objects	55344	6923	62267
# sound objects	56158	7666	63824
# g-vQA boxes	6795	2579	9374

Table A1: Object tracks involved in actions, sounds, and grounded-vQA, split by camera motion.

Camera	**Static**	**Moving**	**Total**
# total points	7791	783	8574
# moving points	3800	783	4583
# static points	3991	0	3991

Table A2: Point tracks available in the *Perception Test*, split by point and camera motion.

3 Point Tracking Baselines

The point tracking algorithms take an input query point which are sampled at the first frame when the point is visible, and track only into the future in an online setting. During evaluation, we ignore the predictions for frames earlier than the query frame, as the algorithm can easily assume the previous points are occluded. We report two baseline results based on this setup: (1) a static dummy baseline assuming all future points are visible and never change the location, (2) a TAP-Net model [19] trained on Kubric and transferred zero-shot.

Following [19], we use three evaluation metrics. (1) *Position Accuracy ($< \delta^x$)*: for a given threshold δ, we measure the fraction of points that are within the threshold of their ground truth, for frames where points are visible. For all predictions, we resize them to 256x256 resolution and measure $< \delta^x$ across 5 thresholds: 1,2,4,8, and 16 pixels. (2) *Occlusion Accuracy (OA)*: a simple classification accuracy for the point occlusion prediction on each frame. (3) *Jaccard at δ*: an evaluation metric considering both occlusion and position accuracy. It is the fraction of 'true positives', i.e., points within the threshold of any visible ground truth points, divided by 'true positives' plus 'false positives' (points that are predicted visible, but the ground truth is either occluded or farther than the threshold) plus 'false negatives' (groundtruth visible points that are predicted as occluded or the prediction is farther than the threshold). Our final metric *Average Jaccard (AJ)* averages Jaccard across all 5 thresholds: 1,2,4,8, and 16 pixels.

Table 3 shows the evaluation results for point tracking based on the three metrics. To further understand the performance, we split points into two groups: static and moving. Note that there are no static points in the moving camera scenario, all points are moving. In static camera, we determine a point is moving if its distance between start frame and end frame is more than 0.01 in the normalized image coordinate system. As expected, the dummy baseline performs well on static points, reaching 0.722 average jaccard. But TapNet significantly outperforms when points are moving, particularly in the moving camera setup, improving average jaccard from 0.088 to 0.328. Besides AJ, TapNet significantly improves the static baseline on occlusion accuracy from 0.675 to 0.849. One interesting observation is that on both position accuracy ($< \delta^x$) and jaccard at δ, TapNet starts to outperform static baseline only when measured above 4 pixel threshold. This is because human annotations still contain small localization errors and 4 pixel threshold is more reliable than 1 or 2 pixel threshold for measuring under 256x256 resolution.

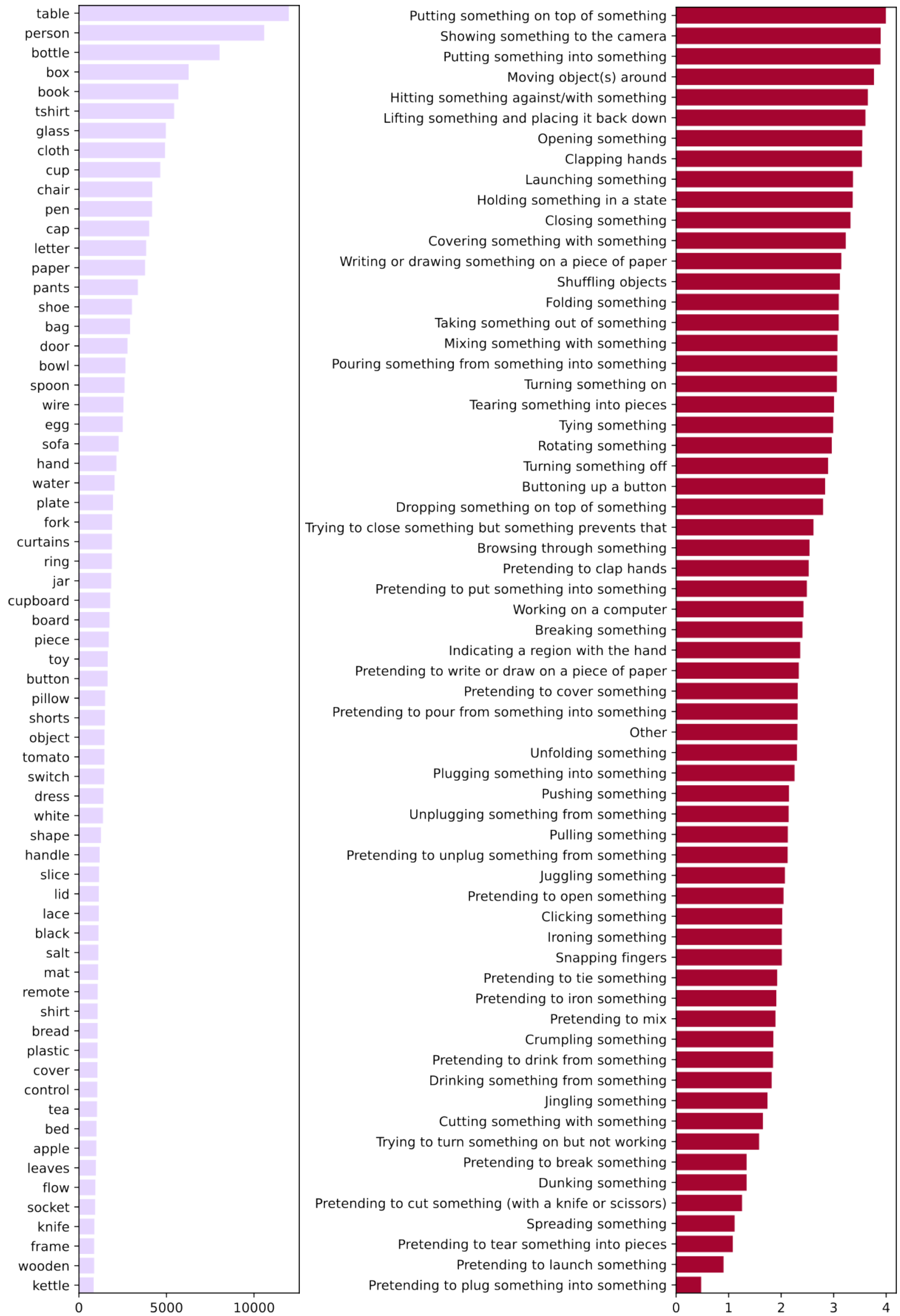

Figure A2: Frequency of objects and log-scale frequency of actions in the *Perception Test*.

4 Temporal action and sound localisation baselines

For the temporal action and sound localisation baselines, we use features extracted with pre-trained encoders. For video, we use TSP features extracted using a Resnet(2+1)D-34 model pretrained on ActivityNet [5]. The resulting features have 512-dim and an effective stride of 32 (corresponding

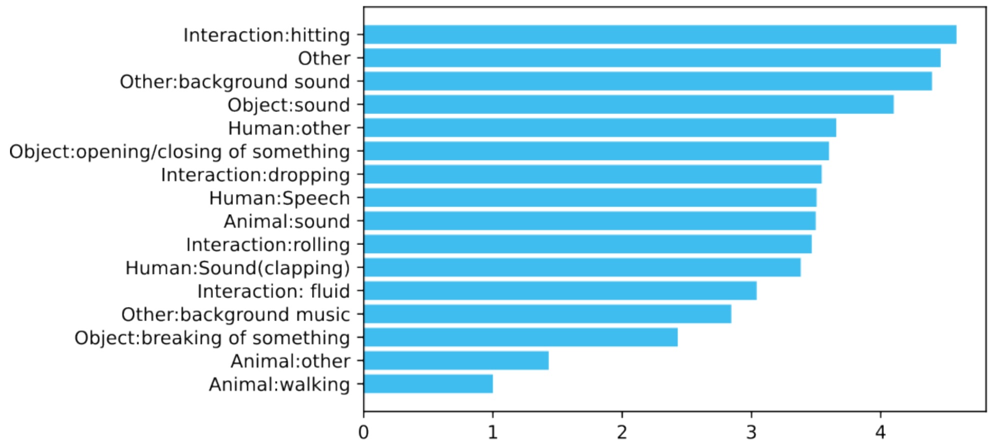

Figure A3: Log-scale frequency of sounds in the *Perception Test*.

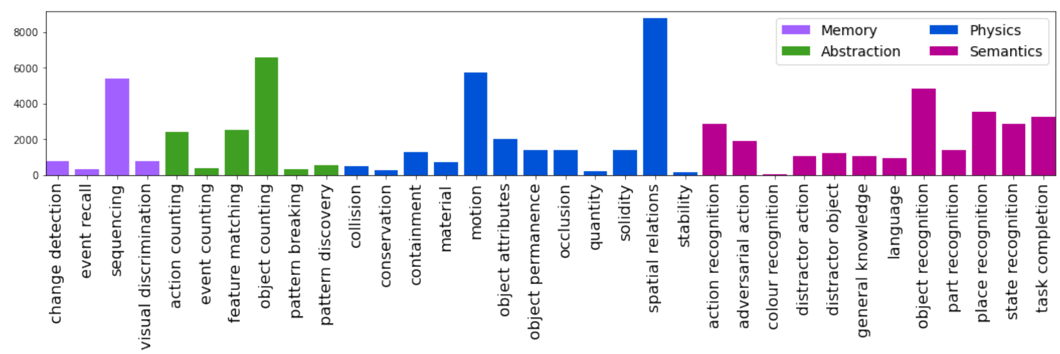

Figure A4: Number of multiple-choice video question-answers in the *Perception Test* across skills in the four skill areas: Memory, Abstraction, Physics, Semantics. One skill can be assigned to multiple skill areas—here we choose one as the prime area for each skill.

roughly to one feature per second): every other input frame is skipped and the model performs a temporal downsampling of 16.

For audio, we extract features using the S3D model pre-trained on AudioSet from MMV [2], with window length of 960ms, window stride 16000. The input audio is downsampled from 48khz to 16khz (keeping every third sample). This results in roughly 2 features per second, each of dimension 256. When using multimodal inputs, the video features are tiled over time (factor 2) to align them with the audio features.

Figure A5 shows the confusion matrix for the action localisation task, normalised by columns. It can be observed that the less frequent actions are often confused with more frequent ones and the model also confuses pretend actions with their non-pretend versions, *e.g. ironing something* vs *pretending to iron something* or *writing or drawing something* vs *pretending to write or draw*.

5 Grounded videoQA baseline

Figure A6 shows HOTA metrics, which integrate detection, association, and localisation scores for the grounded videoQA baseline formed of MDETR detector and Stark tracker.

Point tracking	static points static camera	moving points static camera	moving points moving camera
static baseline	0.722	0.373	0.088
TapNet [19]	0.496	0.399	0.328

Point tracking	OA	$< \delta^0$	$< \delta^1$	$< \delta^2$	$< \delta^3$	$< \delta^4$
static baseline	0.675	0.395	0.512	0.601	0.695	0.784
TapNet [19]	0.849	0.055	0.214	0.687	0.927	0.956

Point tracking	Jac. δ^0	Jac. δ^1	Jac. δ^2	Jac. δ^3	Jac. δ^4
static baseline	0.217	0.301	0.364	0.429	0.495
TapNet [19]	0.025	0.104	0.442	0.699	0.734

Table A3: Static baseline vs TapNet results on the validation set. **Top**: Average Jaccard (AJ), higher is better. The static and moving points are based on the point motion described in Appendix 3. There are no static points in the moving camera scenario. **Middle**: Occlusion Accuracy (OA) and Position Accuracy ($< \delta^x$), higher is better. TapNet outperforms static baseline when measured above 4 pixel threshold. **Bottom**: Jaccard at δ, higher is better. TapNet outperforms static baseline when measured above 4 pixel threshold.

6 Dataset Splits Generation

The 11.6k videos in the *Perception Test* are split into train, validation, and held-out test splits each with roughly $20\%/50\%/30\%$ of the videos respectively. These splits were generated by respecting two constraints: (1) all videos from each unique combination of (`script_id`, `participant_id`) are kept in the same split; more specifically, each script was filmed by a given participant possibly with multiple camera configurations, *e.g.* from different viewpoints, or both with static and moving cameras. The above constraint ensures that all such variations of a script shot by a participant belong in the same split to avoid any leakage of video content across splits, and (2) various video attributes (camera motion, indoor *vs.* outdoor) and annotations are divided in the same proportion across splits, *e.g.* each split will have approximately the above specified fraction of videos with moving camera, or with point annotations. In particular, each question in the multiple-choice and grounded video QA tasks applies to a number of videos; this constraint ensures that these videos are distributed across splits in the specified proportion, such that all questions are present in all the splits.

The above was executed by setting up a linear program with a binary decision variable for each unique (`script_id`, `participant_id`) pair indicating which of the two splits it should be assigned to, denoted collectively $\mathbf{x}\in\{0,1\}^n$ with n being the number of such unique pairs. Note for splitting into three splits, the problem is solved twice sequentially. A feature count matrix $A\in\mathbb{R}^{n\times d}$ was constructed, with A_{ij} being the number of videos shot by the i^{th} (`script_id`, `participant_id`) having the j^{th} video-attribute (d being the total number of video attributes). An "attribute" indicating the total number of videos with a given (`script_id`, `participant_id`) was also included to enforce the number of videos in each split. The following linear program was solved using the CVXPY interface to the MOSEK mixed-integer solver.

$$
\begin{aligned}
\min_{\mathbf{x}} & \left[\left(\max_i \, (1-t_i)^2 \right) + \frac{1}{d} \sum_{i=1}^{d} (1-t_i)^2 \right] \\
\text{s.t.,} \quad & t_i = \frac{A_i^T \mathbf{x}}{\lceil f_1 A_i^T \mathbb{1} \rceil}, \forall i \in \{1, \dots, d\} \\
& (1-\lambda) \leq t_i \leq (1+\lambda), \forall i \in \{1, \dots, d\} \\
\text{and,} \quad & \mathbf{x}_j \in \{0,1\}, \forall j \in \{1, \dots, n\}
\end{aligned}
$$

with A_i being the i^{th} column of A, $f_1 \in [0,1]$ being the target fraction for the split corresponding to label $\mathbf{x}_j = 1$ (*e.g.* $f_1 = 0.5$ for a 50% test split), and $\lambda = 0.25$ is the maximum allowed fractional deviation from the target value. There were $n = 7288$ unique (`script_id`, `participant_id`) pairs, and $d = 249$ video attributes.

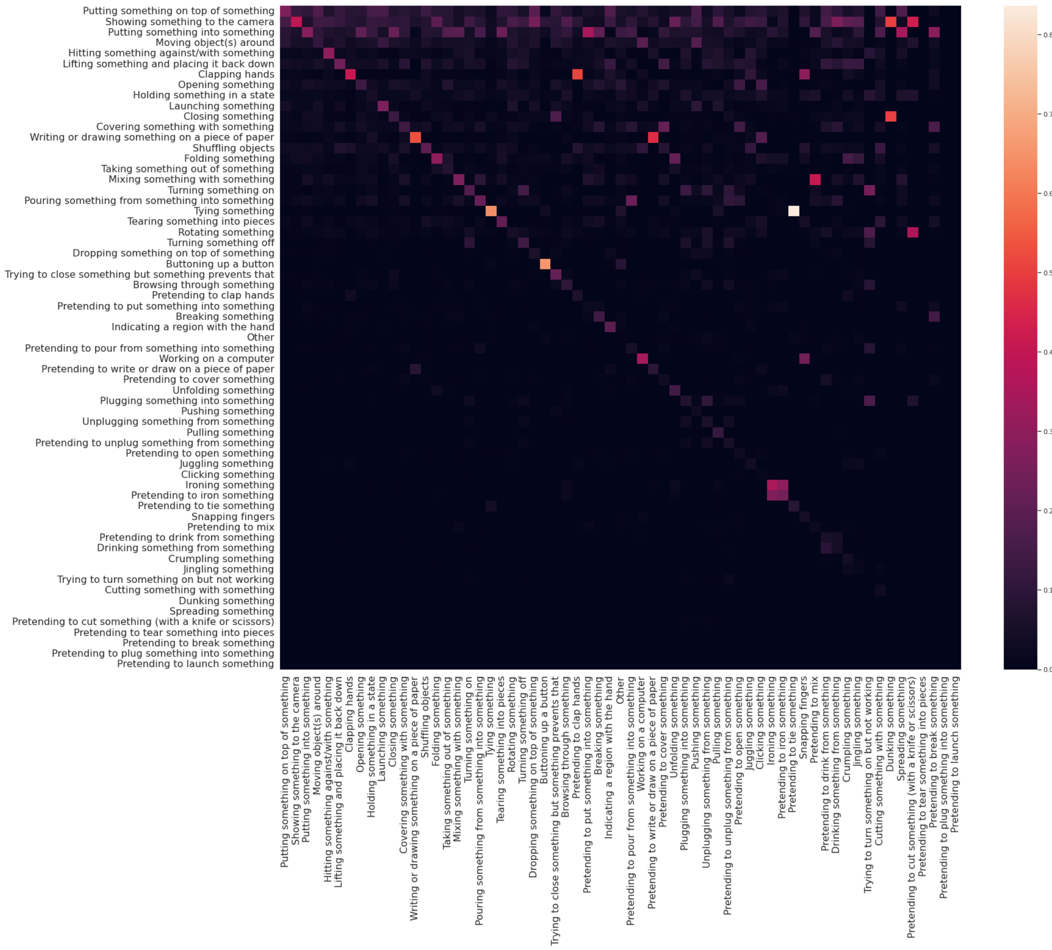

Figure A5: Confusion matrix for ActionFormer predictions on the action localisation task. To be considered as a prediction for a certain segment, the model's confidence has to be above 0.1 and IoU threshold between the prediction and ground truth above 0.1. Ground truth actions are listed on the y-axis, sorted by their frequency; entries are normalised by rows. The less frequent actions are often confused with more frequent actions. The model also confuses pretend actions with their non-pretend versions, *e.g. ironing something* vs *pretending to iron something* or *writing or drawing something* vs *pretending to write or draw*.

7 Annotation collection and cleaning

The different types of annotations were collected using two different approaches:

1. *sequential pipeline* for the object and point tracks, action and sound segments: (i) a rater annotates a video for a given task, (ii) a second rater checks the annotation, makes any necessary corrections, then marks the annotation as complete; (iii) a third rater checks if the annotation is indeed complete or it needs additional changes, in which case they will send the video back to step (ii) to be reviewed by a different rater. For difficult tasks like point tracking or object tracking with hard occlusions, we did multiple annotation cleaning rounds, each time with specific cleaning guidelines. For example, for the videos in cups-games category mentioned above, in one cleaning round, the raters were asked to pay attention to the hidden object, or for videos where the person shows objects to the camera sometimes repeating the same object, we asked raters to pay attention to assign the same object ID when the object reappears. Having videos grouped by script type helped in designing specific cleaning guidelines to ensure good annotation quality.

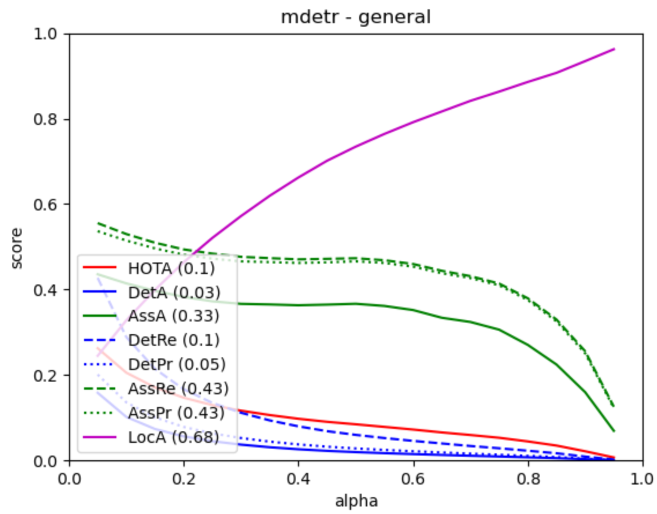

Figure A6: HOTA metrics for MDETR+Stark tracker baseline on the validation split of the *Perception Test*.

2. *parallel pipeline* for multiple-choice and grounded videoQA: multiple raters answer in parallel the same question for the same video and the option chosen by the majority of raters is kept as final answer. Note that for multiple choice QA, during annotation collection, the raters were presented with more than 3 options in some cases. For the final dataset, as the goal was to have the same number of options for all the questions, we chose to keep 3 options to accommodate binary questions as well (where the options used are: *Yes*, *No*, *I don't know*). For questions with more than 5 options, the negative options were sampled based on their frequency as correct options for videos in the same script type. Finally, for some generic questions, *e.g. Which statement describes the scene better?*, the answers were collected initially in open-language format, and then negatives were sampled using the answers from other videos in the same script type, with additional checks from the research team to avoid ambiguous distractors.

As a sanity check, for the action and sound annotations, we checked for overlapping objects involved in both action and sounds (see Figure A7). We observed strong correlations across pairs of action-sound, indicating consistent annotations across modalities, *e.g.* the *Pouring something into something* action shares the same objects with the *Interaction: Fluid* sound, the *Clapping hands* action co-occurs with the *Human (clapping)* sound, the *Lifting something and putting it back down* action co-occurs with the *Object: Hitting* sound, *Moving something around* actions co-occurs with *Object: Rolling* sound, and so on.

8 Diversity of participants involved in filming

To have good visual diversity in the dataset, we selected participants from different countries, having different ethnicity and gender. We include in Table A4 and Figure A8 the self-reported demographics.

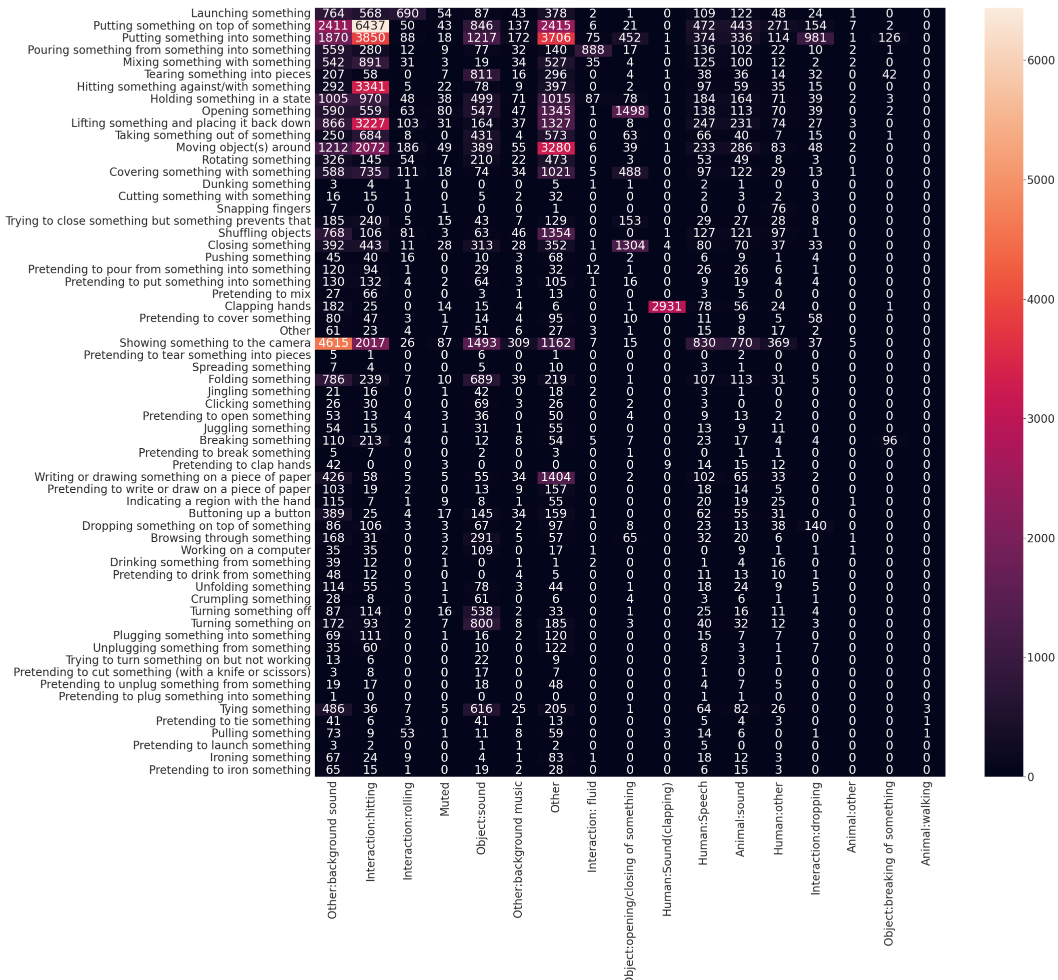

Figure A7: Correlation between action and sound temporal annotations in the *Perception Test*.

Gender	%
Male	46.40
Female, Other	53.60
Ethnicity	**%**
White or Caucasian	28.97
South and East Asian	25.49
Black or African American	21.68
Latino or Hispanic	9.25
Mixed	3.94
Middle Eastern	3.37
Other	7.30

Country	%
Philippines	31.38
Brazil	11.27
Kenya	10.02
Indonesia	8.75
Italy	8.03
Romania	7.57
South Africa	5.25
Turkey	4.12
India	3.72
Mexico	1.45
Bulgaria	1.37
United States	0.70
Egypt	0.48
Other	5.87

Table A4: Self-reported demographics (Gender, Ethnicity, Country) of participants involved in filming.

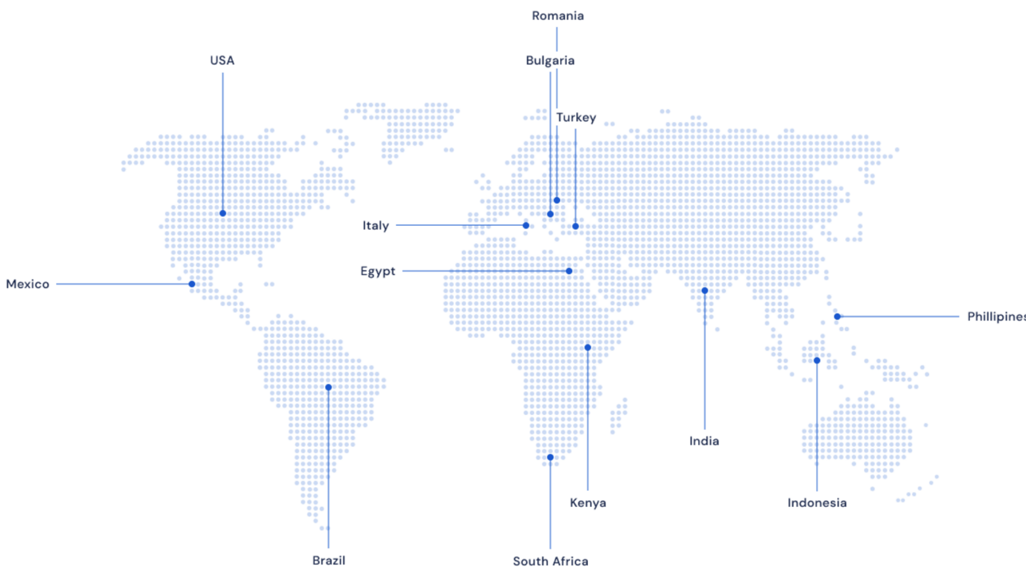

Figure A8: Geolocation of participants involved in filming.