Human vs. Machine Minds: Ego-Centric Action Recognition Compared

Sadegh Rahmaniboldaji*
University of Surrey
United Kingdom
S.Rahmani@surrey.ac.uk

Filip Rybansky*
University of Newcastle
United Kingdom
F.Rybansky2@newcastle.ac.uk

Quoc Vuong
University of Newcastle
United Kingdom
Quoc.Vuong@newcastle.ac.uk

Frank Guerin
University of Surrey
United Kingdom
F.Guerin@surrey.ac.uk

Andrew Gilbert
University of Surrey
United Kingdom
A.Gilbert@surrey.ac.uk

Abstract

Humans reliably surpass the performance of the most advanced AI models in action recognition, especially in real-world scenarios with low resolution, occlusions, and visual clutter. These models are somewhat similar to humans in using architecture that allows hierarchical feature extraction. However, they prioritise different features, leading to notable differences in their recognition. This study investigated these differences by introducing Epic ReduAct [1], a dataset derived from Epic-Kitchens-100. It consists of Easy and Hard ego-centric videos across various action classes. Critically, our dataset incorporates the concepts of Minimal Recognisable Configuration (MIRC) and sub-MIRC derived by progressively reducing the spatial content of the action videos across multiple stages. This enables a controlled evaluation of recognition difficulty for humans and AI models. This study examines the fundamental differences between human and AI recognition processes. While humans, unlike AI models, demonstrate proficiency in recognising hard videos, they experience a sharp decline in recognition ability as visual information is reduced, ultimately reaching a threshold beyond which recognition is no longer possible. In contrast, the AI models examined in this study appeared to exhibit greater resilience within this specific context, with recognition confidence decreasing gradually or, in some cases, even increasing at later reduction stages. These findings suggest that the limitations observed in human recognition do not directly translate to AI models, highlighting the distinct nature of their processing mechanisms.

1. Introduction

Ego-centric action recognition is the process of identifying and interpreting human actions from a first-person perspective, playing a crucial role in both human cognition and artificial intelligence (AI) models. Ego-centric action recognition has essential applications for assistive technology, augmented reality, human-computer interaction, and wearable AI systems. Understanding actions from an ego-centric viewpoint could enable applications like hands-free device control, activity monitoring, and robotic assistance.

In humans, action recognition relies on two parallel pathways: the ventral "what" and the dorsal "where and how" pathway [26, 27, 66]. These are hierarchically specialised for different information, such as form and motion. Similarly, earlier AI models like Convolutional Neural Networks (CNNs) [41, 42] use a hierarchical structure but cannot perceive different information. Other AI models, such as Two-Stream Networks [67], LSTMs [33], and Vision Transformers (ViTs) [17] emulate human action recognition by integrating spatial and temporal information. However, they struggle with occlusions and lack predictive processing capabilities, which allow humans to infer missing details [19].

The architecture of many AI visual recognition models is constructed into hierarchical layers. Researchers have shown that activation in these layers can predict neural responses in early [29], intermediate [75], and late [74] areas of the monkey and human visual cortex. Thus, some researchers argue that their performance and congruity with behavioural and neural responses make them adequate models of human recognition [12, 44], but others counter that they offer only slightly above chance consistency with humans [24, 25].

Despite similarities in hierarchical architecture, the features extracted at each layer in human vision and AI models may differ. For example, humans initially process boundaries and surfaces, potentially leading to a shape bias [45]. AI models, on the other hand, often show a texture bias

*These authors contributed equally.

[1]The dataset is available at: https://github.com/SadeghRahmaniB/Epic-ReduAct

[22, 23, 48]. AI models can also fail to encode the global arrangement of visual elements [6], three-dimensional and internal structure of objects, and occlusion and depth information [4, 12, 16, 37, 73]. However, specialised training datasets or architecture can produce AI models that extract features more closely aligned with humans [22, 38, 56].

While human vision and artificial models use hierarchical representations, humans uniquely integrate bottom-up sensory input and top-down cognitive processes for dynamic perception [1]. Human perception is continually refined by prior knowledge at inference time [7, 34], whereas deep learning models predominantly depend on labeled data at training time. Moreover, human attention naturally prioritises salient regions based on contextual relevance [71].

Given these discussions, this paper investigates how humans and AI models differ in their ability to recognise activities from video segments in challenging ego-centric scenarios. Ego-centric videos provide enhanced information about objects within an actor's immediate affordance [11, 60] and their interactions via hands [13], yet they often offer less contextual information [32]. This reduced context may influence both the visual features [43, 50] and cognitive strategies [54, 59] employed during action recognition. Consequently, ego-centric video analysis represents an ideal domain for comparative investigation. We used videos from the standard ego-centric kitchen activity dataset, Epic-Kitchens-100 [15], and systematically reduced the available spatial and temporal information, creating the **Epic ReduAct** (Epic-Kitchens Reduced Action Videos) dataset, which will be publicly released. To quantify human and AI performance, and drawing on paradigms used for static images [9] and third-person actions [10], we adopted the idea of Minimal Recognisable Configurations (MIRCs)—the smallest spatial crops of images still identifiable by humans, with smaller quandrants, known as sub-MIRCs, unrecognisable. Using Epic ReduAct and the MIRC definition, we quantified performance disparities between humans and AI models through two distinct metrics: the newly introduced **Average Reduction Rate** and a second metric inspired by Ben-Yosef's work [9], the **Recognition Gap**. These metrics provide insights into the key distinctions between human and model-based action recognition, highlighting potential avenues for enhancing AI models. The main contributions of this paper are, therefore:

1. **Creation of Epic ReduAct dataset:** Systematically reduced ego-centric videos (in spatial and temporal resolution) from the Epic-Kitchens-100 dataset, generating a publicly available benchmark dataset to study minimal requirements for ego-centric human action recognition.
2. **Novel Evaluation Metric:** Introduced a quantitative metric, Average Reduction Rate, to systematically quantify and interpret differences in recognition capabilities between humans and AI models.
3. **Human vs. AI Action Recognition Analysis:** We conducted comparative experiments between over 3000 human participants and a state-of-the-art AI video understanding model using the concepts of MIRC and sub-MIRC to uncover significant differences in how each recognises actions from reduced video information.

2. Related Works

2.1. Human and Computer Vision

Human action recognition relies on two parallel neural pathways: the ventral and dorsal streams [26, 27, 66]. These pathways originate from the primary visual cortex (V1) and extend through higher-level areas such as the inferotemporal (IT) and prefrontal cortex (PFC) [65]. Initial visual processing extracts basic features like edges, colour, and motion [36, 61, 62], which are progressively integrated into more complex representations. The reverse hierarchy theory [1] suggests rapid bottom-up categorization is followed by top-down refinement, enabling precise action recognition [7, 8, 34]. Inspired by these mechanisms, AI models incorporating motion-based attention have achieved state-of-the-art results in video action classification [2].

In contrast, AI-based action recognition relies on deep learning models trained on large datasets. Early methods employed convolutional networks such as AlexNet [41], ResNet [31], and VGG [68], designed for static image analysis. Despite their layered structure inspired by human vision, their performance remained low. Recent advances, including Transformers [72], Vision Transformer [17], and contrastive models like CLIP [55], have enhanced visual understanding. Followed by video specific architectures, such as Side4Video [76] and MOFO [2], leveraging motion-based self-supervised learning, while FILS [3] integrates language semantics. Through incorporating more Multimodal data, more recent Large language models (LLMs) and Video language models (VLMs) like VideoLlama3 [77] and VideoChatGPT [47] allow for a deeper scene interpretation. Comparing human and computer vision offers insights into improving AI models and understanding human perception. While humans excel at generalization and contextual reasoning, AI systems achieve high accuracy but often lack adaptability. Bridging these gaps through biologically inspired architectures and multimodal learning could lead to more robust artificial vision systems.

2.2. Biologically-inspired Vision Models

Early neuro-computational models drew inspiration from the hierarchical organisation of human visual pathways. A pioneering biologically-inspired model for visual recognition, the HMAX model [39, 58, 65], processes retinal output through a neurophysiologically consistent hierarchy of alternating simple and complex cell layers using linear or nonlinear MAX pooling operations. Giese et al. [26] extended this framework to action recognition tasks (e.g., walking, running), employing a two-pathway feedforward architecture modeled after the human ventral and dorsal pathways. Researchers have recently proposed methods to improve

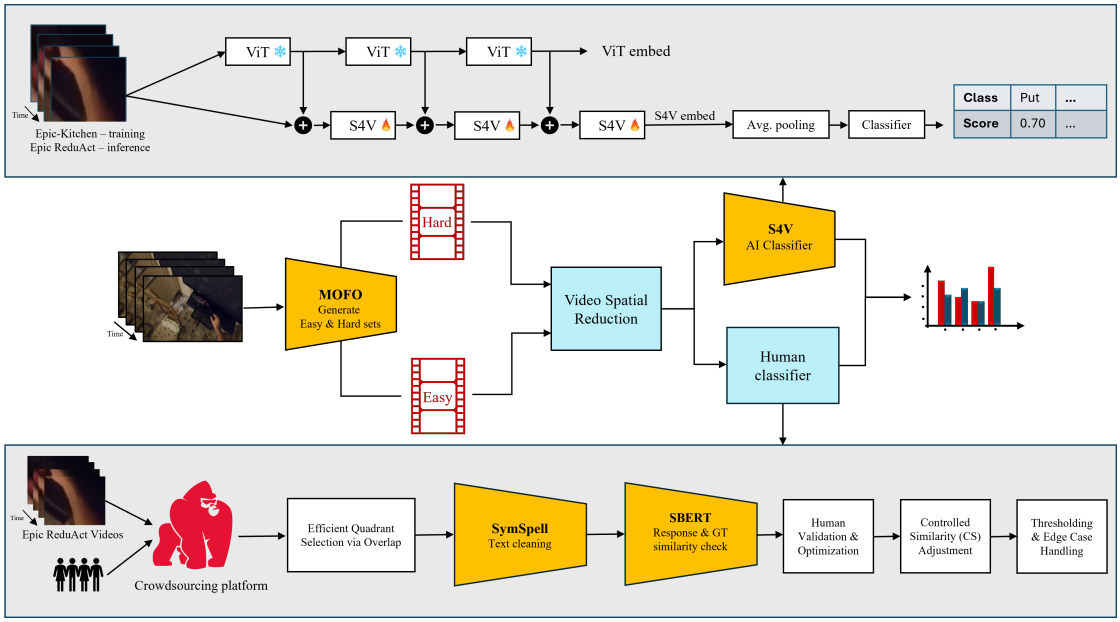


Figure 1. Research pipeline highlighting different processes, including data preparation and classification by human and AI classifiers

the HMAX models within biological constraints. These include the incorporation of layers corresponding to the full anatomy of the visual cortex [64], modifying the sequence of information flow and weights according to context [40, 49] or temporal coordination of neural activity [51]. Despite their valuable insights into biological visual recognition processes, these models typically do not match the high-level performance of recent vision models described earlier.

2.3. Joint AI and Human Studies

Research explicitly integrating human-inspired methods into computer vision has explored informative image regions to improve recognition, particularly in partially occluded scenarios [20], as well as aligning receptive fields of neural networks more closely with human vision [19]. Studies have suggested aligning AI model behaviour with human perception, such that image alterations unnoticed by humans should similarly not affect model predictions [35]. Comparative studies have consistently demonstrated differences between human and AI models in their visual attention and recognition performance for image classification tasks [14, 18, 28, 30, 53, 70]. Notably, [9] introduced the concept of Minimal Recognisable Configurations (MIRCs) and sub-MIRCs in static images, revealing that human recognition accuracy sharply declines upon removing critical features. In contrast, AI models experience a more gradual accuracy reduction. Extending this approach to videos, [10] identified critical spatial and temporal configurations necessary for human action recognition in third-person action videos from the UCF101 dataset [69]. They found a significantly more significant recognition drop-off from MIRCs to spatial sub-MIRCs in humans than in the model. However, there were no human-model differences in the drop-off for temporal sub-MIRCs. Our research builds upon this comparative framework by exploring ego-centric video, a domain previously not extensively studied, and by examining a broader range of ego-centric activities.

3. Method

Our research pipeline, illustrated in Fig. 1, outlines our approach to comparing human and AI performance in ego-centric video action recognition. We began by employing a classifier to pre-select easy and hard video sets. To enable a comparison between how human and AI models recognise activities in video, we artificially and systematically reduced the video's spatial resolution. Then, using human participants and an AI model as classifiers, we evaluate and compare human and AI performance on these spatially reduced videos to quantify the difference in recognition between the human and AI models.

3.1. Problem Definition and Epic ReduAct Preparation

To enable this investigation, we first created an *Easy* and *Hard* subsets of the Epic-Kitchens dataset [15] that represents different levels of activity recognition difficulty for AI models. Each set comprises of 18 Epic-Kitchens videos with a mean duration of 2.35s (Standard Deviation (SD) duration = 1.11s), to enable comparisons between human and AI model performance on distinct difficulty levels.

To generate the Easy and Hard sets, we employed a state-of-the-art AI model [2] to predict classification probabilities for various action classes. Videos with top-1 prediction

Catergory	Videos	Samples	MIRCs	sub-MIRCs	Verb Classes
Easy	18	4,503	273	1092	close, cut, hang, open, pour, put, remove, take, turn-off, turn-on, wash
Hard	18	3,173	402	1608	close, hang, insert, open, peel, pour, put, remove, serve, take, turn-off, wash

Table 1. Epic ReduAct dataset details

confidences exceeding 60% were identified as candidate Easy videos. Conversely, videos where the correct label failed to appear among the top 5 predictions were marked as candidate Hard videos. A subsequent manual review eliminated ambiguous or overly similar samples. The final basis of Epic ReduAct includes actions from 11 verb classes for the Easy set and 12 verb classes for the Hard set.

Next, we conducted online experiments to systematically reduce the spatial information of the 18 Easy and 18 Hard videos (36 total) across eight hierarchical levels to identify MIRCs. The process is shown in Fig. 2 for a video with the GT label *close*. At Level 0, we spatially cropped a region of the video that best encompassed the entirety of each video. At Level 1, frames from each parent video were cropped at the four corners, generating four child sub-videos per original. Levels 2 through 7 involved recursively applying this corner-cropping method to each subsequent generation of parent videos.

The labelling of a video as a MIRC and sub-MIRC [9] was determined as follows. After spatial cropping of a parent video into four quadrants, quadrants recognisable by at least 50% of human participants underwent further recursive reduction. Conversely, if none of the quadrant sub-videos from a parent were recognised by more than 50% of the participants, the parent video was classified as a **MIRC** and the unrecognisable child quadrants as **sub-MIRCs**.

Given the exponential increase in child quadrants at the higher levels, we implemented quadrant pruning to select only the most informative quadrants for testing. This helped identify MIRCs more efficiently and quickly. The quadrant pruning occurred as follows:

1. After testing a Level of quadrants described later in Sec. 3.2.2, we produced child quadrants by cropping parent quadrants that were recognised.
2. We then assumed that child quadrants, whose parent quadrant was recognised but fully contained within a different unrecognised quadrant from any previous level, would also not be recognised. Therefore, we did not test those at the next level.
3. We also assumed that child quadrants from a single Level that overlapped each other by at least 95% would have equal accuracy. Therefore, we tested only one quadrant from each such cluster.
4. From the remaining quadrants in the level, we selected child quadrants less likely to be recognised (defined as overlapping a not recognised quadrant from any previous level by 65% or more) and those most informative (defined as having the highest cumulative surface area overlap with other child quadrants from the same level), until we filled the maximum number of quadrants per video that could be efficiently tested at that level.

This procedure generated 7,676 videos, including the Level 0 videos. The final Epic ReduAct dataset is summarised in Tab. 1.

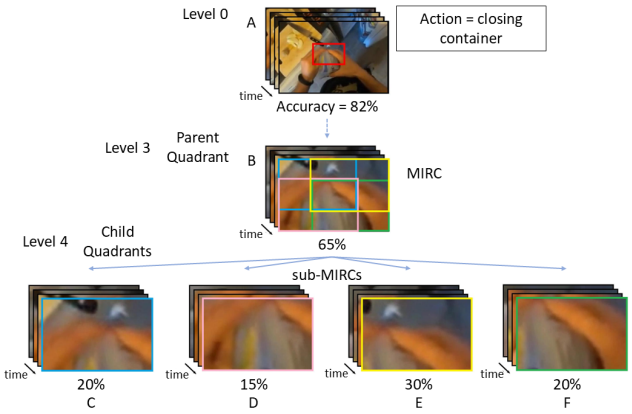


Figure 2. Example of applying reductions to a video. Video GT was *close*. A was cropped for 3 Levels before becoming Quadrant B recognised by 65% of participants. B was the Parent Quadrant to 4 Child Quadrants (C–F), each recognised by fewer than 50%, making them sub-MIRCs and B the MIRC.

3.2. Classifiers

Given the spatially reduced videos, two recognition classifiers, human observers and an AI model, were used to compute and identify the MIRCs and sub-MIRCs. Before introducing the classifiers, it is essential to define accuracy within the context of both models. In the AI model, accuracy is the confidence score of the predicted verb relative to other verbs, represented as a fraction of 1 (i.e. this is simply the confidence output by the model after softmax). Similarly, accuracy is defined as the proportion of individuals who correctly labelled the video for human participants.

3.2.1. AI Classifiers

The AI classifier used in this study was based on the Side4Video (S4V) framework [76], ensuring the hard and easy sets were unbiased by the MOFO-based selection [2],

which uses a fundamentally different, unsupervised training process. Moreover, while the MOFO performance was comparable to that of S4V, the latter demonstrated substantially faster training times, rendering it a more practical and efficient choice for integration into our pipeline. As shown in the top part of Fig. 1, S4V employs the vision module of OpenCLIP [55] as a video feature encoder. It introduces a lightweight spatial-temporal side network attached to a frozen pre-trained vision model. The trainable layers are integrated in parallel to the larger frozen model layers rather than sequentially. This allows efficient fine-tuning for video understanding tasks without back-propagating through the large pre-trained model, leveraging the extensive multi-level spatial features from the original image model while significantly reducing memory usage—up to 75% compared to previous adapter-based methods—and facilitating the effective transfer of substantial models such as ViT-E (4.4B parameters) to video tasks. The AI model was trained on the training split comprising 67179 videos from the Epic-Kitchens dataset [15], excluding the 36 videos selected for comparison with human performance. Each input video clip was subsampled to 8 frames with a spatial resolution of 224×224, and frames were uniformly spaced. The video model uses ViT-B/16 [17] and was trained using the AdamW optimiser [46]. This trained classifier was subsequently applied to spatially reduced videos in Epic ReduAct to evaluate recognition performance against the ground-truth verbs for the original videos, with the results discussed in Sec. 4. During inference, the study prioritises recognising ongoing actions over object identification. Accordingly, verb accuracies in predicted labels were evaluated by aggregating confidence scores for all correctly predicted verbs within action labels.

3.2.2. Human Classifiers

For the human classifiers, experiments were set up on the Gorilla platform [5] and disseminated via Prolific (www.prolific.com).The sample involved responses from 3800 participants (1964 females; 2329 males; and 54 others; mean age = 33.2 years, SD age = 11.3 years). Based on the size of the assigned dataset, median completion time and participant remuneration ranged from 9 to 21 minutes and £0.70 to £2.75 for fair reimbursement. Ethical approval was obtained from the university's Research Ethics Committee (Ref: 38465/2023). The experiment began with five practice trials. Participants were then tested with all 36 videos (18 easy and 18 hard) to measure recognition accuracy, where each participant was randomly assigned to view only one child quadrant per video, intermixed with two catch trials. The practice and catch trials used additional easy videos from the EPIC-Kitchens dataset [15]. Seventeen participants were replaced due to poor performance on catch trials. Each trial began with a fixation cross displayed in the centre of the screen for 500ms. The video was presented as centred on a white background until a response was recorded. After 4000ms, the participant was asked to type their response about what action was being performed inside the box. Humans have been asked to type their answer to avoid being biased into pre-selected labels. The participant identified a single action followed by a single object being acted upon, using up to three words.

The correctness of participants' responses was assessed using an optimised controlled similarity (CS) measure to a Ground Truth (GT) label for each video. Responses were first tidied by removing punctuation, articles and subjects (words like 'man', 'person'), correcting misspellings algorithmically and with the help of SymSpell [21], and manually rewording responses with incorrect word count or slang. We then computed the cosine similarity, S, between the response and GT [63], using the sentence-BERT (SBERT) model "all-mpnet-base-v2" [57]. Additionally, we computed the similarity of isolated actions, S_A, and objects, S_B, using Word2Vec [52]. The Controlled Similarity (CS) score was then calculated as:

$$CS = S - (S_B \cdot p)^2 + (S_A \cdot b)^2 \tag{1}$$

where p and b are penalty and bonus constants, respectively, optimised a priori. The response was correct if its CS with the GT was greater than a set threshold.

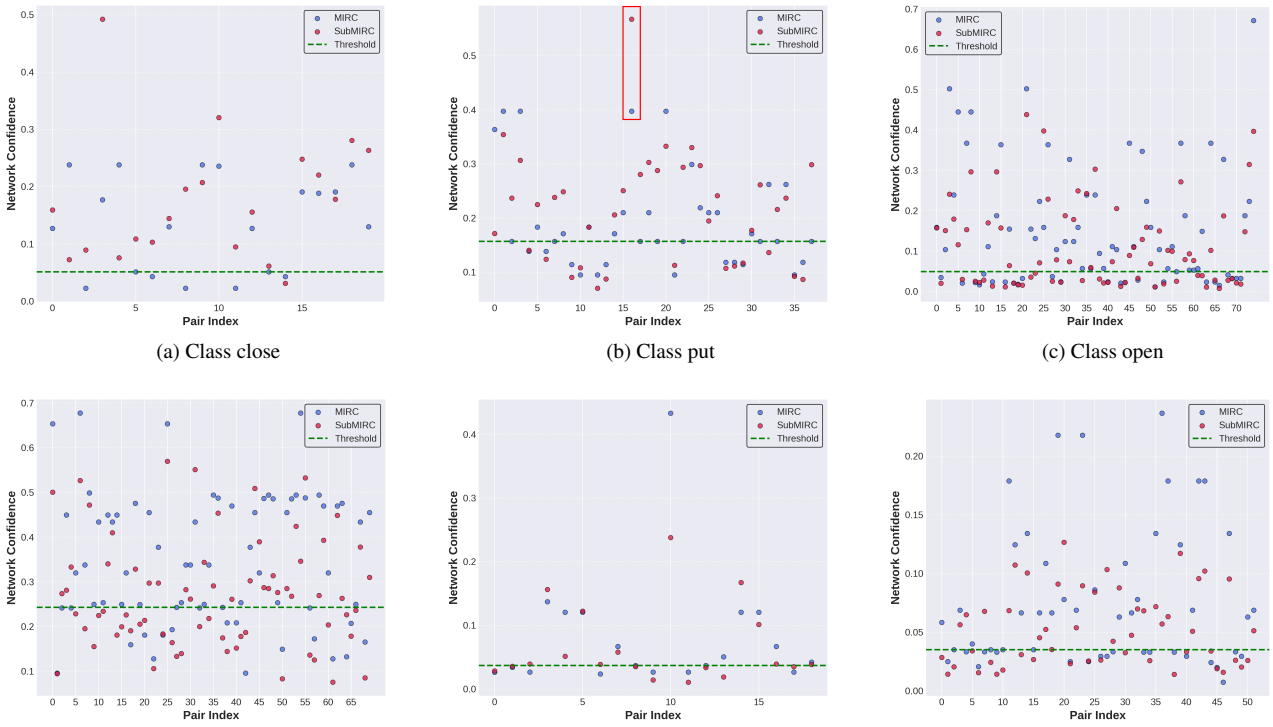
4. Experiments and Results

4.1. Metrics

We employed two approaches to compare the results: Recognition Gap and Average Reduction Rate.

Recognition Gap Metric. For humans, the recognition gap is the difference in average classification accuracies across participants between parent MIRCs and sub-MIRC child quadrants. For an AI model, it is a function of human recognition performance as the human accuracy for MIRCs of a class, X, is the baseline for evaluating AI model predictions. We define a threshold line (tl) as the recognition rate at which AI model predictions exceeded X. This is shown as the green dotted line in Fig. 5. We then applied sub-MIRCs to this threshold and measured the proportion of samples, y, that surpassed it. That class's final recognition gap was $X - y$, representing the adjusted recognition performance relative to human baseline performance. As an example, in Fig. 3d, the MIRC accuracy for humans is 59%, and the threshold line will be the value that the same fraction of MIRCs (59%) identified by the AI model are above that, which for example is a threshold of 0.24 for network confidence for that specific class. Therefore, the recognition gap for that class will be the difference between the average accuracy of MIRCs of that class (59%) and those sub-MIRCs above the threshold line (36%), which is 59% - 36% = 23%.

Average Reduction Rate Metric. The Average Reduction Rate quantifies the decrease in accuracy caused by a spatial reduction, which is the difference between the average accuracy of predicted child videos at a specific level and



(a) Class close (b) Class put (c) Class open

(d) Class take (e) Class wash (f) Class open

Figure 3. Classwise recognition gap of the AI model between the MIRC/subMIRC pair relative to the threshold line, with class-specific acceptance thresholds calibrated to match the human recognition rate. Top row, Fig. 3a, Fig. 3b, and Fig. 3c are three selected classes of easy set which we highlight the sample 16 of class put for further explanations in section 4 (Fig. 4) and the bottom row, Fig. 3d, Fig. 3e, and Fig. 3f illustrate some candidate classes of hard set.

		Classes													
Classifiers	Sets	hang	serve	take	open	remove	turn-off	turn-on	wash	peel	close	cut	put	pour	insert
Human	Easy	+38.65	N/A	+37.46	+37.86	+27.85	+34.18	+31.11	+31.89	N/A	+44.75	+36.59	+36.84	+39.70	N/A
	Hard	+41.70	+42.87	+40.78	+38.36	+37.50	+60.55	N/A	+31.05	+38.20	+33.75	N/A	+34.31	+24.70	+40.00
AI	Easy	+0.02	N/A	-19.25	-5.91	-9.20	-7.36	-16.36	-0.96	N/A	+4.26	-2.68	+4.12	+0.14	N/A
	Hard	-0.04	0.00	-9.00	-3.18	-0.49	-0.21	N/A	-2.08	-0.01	-0.07	N/A	+2.63	0.67	-0.02

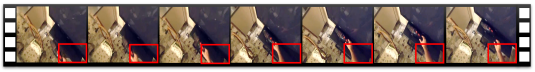
Table 2. The Recognition Gaps for both classifiers, for both sets of data, split by class labels. The values are percentages, and N/A shows that the class is unavailable for that set.

their parents in the previous level. This provides insight into the impact of spatial reduction on the recognition performance of human and AI models. Therefore, it measures cases where a child video becomes less recognisable than its parent, considering only positive reductions.

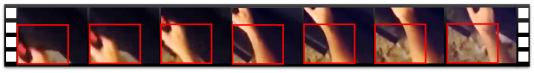
4.2. Recognition Gap

We computed the recognition gap on specific classes in our dataset, with selected example classes shown in Fig. 3 to compare performance between humans and the AI model. These figures illustrate the positioning of each MIRC/sub-MIRC pair concerning the AI model's confidence and the threshold line. Notably, the Figures reveal that the sub-MIRC exhibits higher confidence in many instances than its corresponding MIRC, providing further insight into the negative values observed later in Fig. 5. To illustrate this phenomenon, we presented a video sample that visually demonstrates why, in some instances, the AI model's performance improves despite reduction. Fig. 4 displays a video with the label *put* at three different reduction levels: the original video (no reduction), MIRC (level 2), and sub-MIRC (level 3). A closer examination reveals that the level 2 video contains numerous irrelevant details, such as background elements. In contrast, at level 3, the focus is primarily on the hand. This shift in focus enables the AI model to better recognise the ongoing action, with confidence increasing from 39% at the MIRC level to 56% at the sub-MIRC level. This particular example is depicted by the red (subMIRC) and blue (MIRC) dots in Fig. 3b highlighted with a red rectangle (pair number 16).

This occurs in several other videos and supports the idea that the AI model relies on fine details and textures, whereas humans perceive visual information more holistically. The recognition gaps for all human and AI model classes can be found in Tab. 2. The table shows that the concepts of MIRC and sub-MIRC do not serve as recognition boundaries for AI models like they do for humans. While human recognition performance declines sharply across all classes, AI models generally exhibit improved detection in both the easy and hard sets, as indicated by the predominantly negative recognition gap values. For example, the recognition gap for class *turn-off* for human, easy and hard sets, has reduced by 34.18% and 60.55%, respectively. For the AI model, easy and hard sets, it has improved 7.36% and 0.21%. Notably, the smallest decline in human recognition occurred for the class *pour* in the hard set, with a substantial reduction of 24.70%, highlighting the significant impact of MIRC and sub-MIRC constraints. In contrast, AI models demonstrated a nearly 20% increase in accuracy for the class *take* in the easy set. Even in cases where AI performance declined, such as for the class close in the easy set, the reduction in recognition accuracy was only 4.26%, a marginal change that is not comparable to the drastic declines observed in human recognition gaps.



(a) Original video - No reduction - GT label *put*

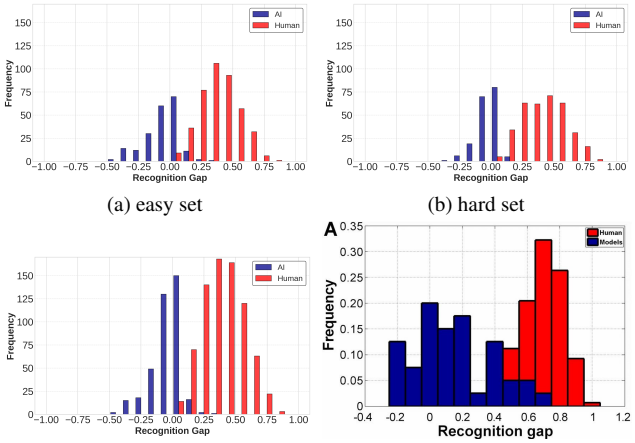


(b) Level 2: The MIRC level

(c) Level 3: The sub-MIRC version of Fig. 4b

Figure 4. Sample of spatial feature loss between original frames, MIRC and sub-MIRC of a video highlighting the cropped areas. The red bounding boxes indicate the location of the spatially reduced child video at the next level.

Fig. 5 presents the recognition-gap frequency distribution for the Easy, Hard and combined sets (Fig. 5a, Fig. 5b, and Fig. 5c), which allows for comparison between humans and AI model. Our results show a similar distribution pattern to previous work with images [9]. Similarly, AI models exhibit some improvement in image recognition, whereas human accuracy consistently declines. Humans also experience a sharper decrease in recognition performance compared to AI models (Fig. 5d). Our results further show that humans are susceptible to substantial losses in recognition confidence. In contrast, spatial reductions can enhance the AI model's ability to detect actions, as evidenced by negative recognition gaps (Fig. 4). The frequency distributions are also broader



(a) easy set (b) hard set

(c) combined (d) From [9]

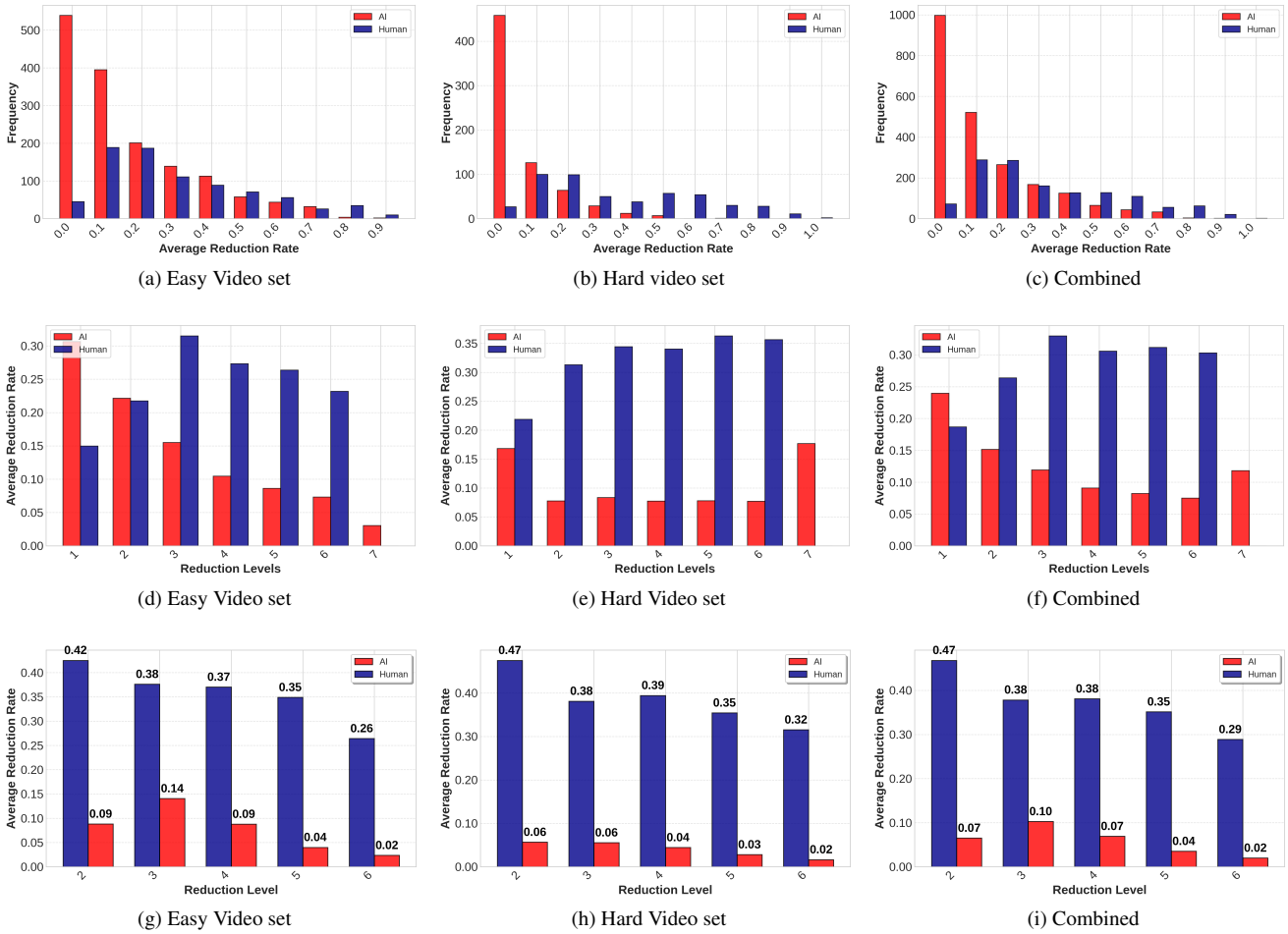
Figure 5. Frequency distribution of recognition gap for AI model and humans on easy and hard video sets. A comparison with the same metric on different data from [9] in Fig. 5d

for humans compared to the AI model, showing more diverse reductions than the AI model reductions in recognition gaps, which are more gradual. These findings indicate that, despite advancements in AI models, the gap between human and machine recognition capabilities persists.

4.3. Average Reduction Rate

This metric, used to compare model performance with psychophysical studies, helps identify levels where significant drops in recognition occur. Fig. 6 compares human classification performance and the results of the AI model, shown through three distinct sets of charts. The first row (Fig. 6a, Fig. 6b, and Fig. 6c) depicts the frequency distribution of the average reduction rate for a parent and child quadrant. It highlights that although human and AI model reduction rates exhibit a similar overall pattern, humans experience a more significant decline in video understanding as indicated by the spread over the larger average reduction rates, indicating more catastrophic recognition failure between the Levels.

The second row (Fig. 6d, Fig. 6e, and Fig. 6f) presents the average reduction rate as a function of the reduction level, providing insight into the specific levels at which the most significant reductions occur. Those charts indicate that the AI model exhibits a more gradual reduction across the levels, suggesting that its accuracy decreases steadily as the spatial information decreases. In contrast, while early reductions have a relatively minor effect on human recognition, recognition performance sharply deteriorates in later levels. This pattern suggests that at specific reduction levels—such as level 3 for easy videos (Fig. 6d)—the degradation of global features significantly disrupts human confidence in accurately detecting actions. The third row (Fig. 6g, Fig. 6h, and Fig. 6i) presents the average reduction rate as a function of the reduction level for only MIRC/sub-MIRC pairs. These



(a) Easy Video set (b) Hard video set (c) Combined

(d) Easy Video set (e) Hard Video set (f) Combined

(g) Easy Video set (h) Hard Video set (i) Combined

Figure 6. Average reduction rates charts. Fig. 6a, Fig. 6b, and Fig. 6c show the distribution of pairs at all levels for every 0.1 increase in the reduction rate. Fig. 6d, Fig. 6e, and Fig. 6f show the average reduction rate as a function of reduction level for all parent-child quadrant pairs. Fig. 6g, Fig. 6h, and Fig. 6i show the same comparison as the average reduction rate per reduction level for only MIRC/sub-MIRC pairs.

charts show a similar trend compared to all parent-child pairs in the second row. A detailed inspection of the differences between the easy and hard sets, including the Recognition Gaps, reveals a significant disparity between human and AI model performance in ego-centric action recognition. Specifically, humans exhibit superior recognition capabilities, as reflected in the higher number of correctly identified videos in Fig. 6b and Fig. 5b compared to Fig. 6a and Fig. 5a. This indicates that humans remain confident in recognising actions in hard videos, whereas AI models struggle to predict the correct labels. This trend can also be interpreted from Fig. 6d, Fig. 6e, Fig. 6g, and Fig. 6h, where the lower number of videos available in the hard set influences the Average Reduction Rate for the AI model.

5. Conclusion

Using Epic ReduAct, a new dataset derived from Epic-Kitchens-100 dataset[15] with systematically reduced spatial and temporal information, this study examined human and AI in action recognition in ego-centric videos. This study highlights key differences between human and AI recognition. Humans excel at so called *hard* videos but struggle as the visual information decreases, eventually reaching a recognition limit. AI models, however, show a greater resilience, with confidence declining more gradually or even increasing at times. These findings suggest that human recognition boundaries do not directly apply to AI models. While this work focused on spatial reduction, future research will explore the role of temporal features in action recognition for humans and AI models.

Acknowledgements

Leverhulme Trust Research Project Grant RPG-2023-079 funded this work.

References

[1] Merav Ahissar and Shaul Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, 8(10):457–464, 2004. 2

[2] Mona Ahmadian, Frank Guerin, and Andrew Gilbert. Mofo: Motion focused self-supervision for video understanding. *NeurIPS 2023 Workshop Self-Supervised Learning: Theory and Practice*, 2023. 2, 3, 4

[3] Mona Ahmadian, Frank Guerin, and Andrew Gilbert. Fils: Self-supervised video feature prediction in semantic language space. *British Machine Vision Conference (BMVC'24)*, 2024. 2

[4] Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4849, 2019. 2

[5] Alexander L. Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jonathan K. Evershed. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52:388–407, 2020. 5

[6] Noah Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14 (12):e1006613, 2018. 2

[7] Moshe Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15(4):600–609, 2003. 2

[8] M. Bar, K. S. Kassam, A. S. Ghuman, J. Boshyan, A. M. Schmid, A. M. Dale, M. S. Hämäläinen, K. Marinkovic, D. L. Schacter, B. R. Rosen, and E. Halgren. Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, 103(2):449–454, 2006. 2

[9] Guy Ben-Yosef, Liav Assif, and Shimon Ullman. Full interpretation of minimal images. *Cognition*, 171:65 – 84, 2018. 2, 3, 4, 7

[10] Guy Ben-Yosef, Gabriel Kreiman, and Shimon Ullman. Minimal videos: Trade-off between spatial and temporal information in human and machine vision. *Cognition*, 201:104263, 2020. 2, 3

[11] Anna M. Borghi, Andrea Flumini, Nikhilesh Natraj, and Lewis A. Wheaton. One hand, two objects: Emergence of affordance in contexts. *Brain and Cognition*, 80(1):64–73, 2012. 2

[12] Jeffrey S. Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E. Hummel, Rachel F. Heaton, and et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023. 1, 2

[13] Francesca Campanella, Giulio Sandini, and Maria Concetta Morrone. Visual information gleaned by observing grasping movement in allocentric and egocentric perspectives. *Proceedings of the Royal Society B: Biological Sciences*, 278 (1715):2142–2149, 2011. 2

[14] Javier Carrasco, Aidan Hogan, and Jorge Pérez. Laconic image classification: Human vs. machine performance. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, page 115–124, New York, NY, USA, 2020. Association for Computing Machinery. 3

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. *European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 5, 8

[16] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. In *NeurIPS'22*, 2022. 2

[17] 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. In *International Conference on Learning Representations*, 2021. 1, 2, 5

[18] Leonard E. van Dyck and Walter R. Gruber. Seeing eye-to-eye?: A comparison of object recognition performance in humans and deep convolutional neural networks under image manipulation. Workingpaper, 2020. 19 pages, 7 figures, 3 tables. 3

[19] Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432–9446, 2022. 1, 3

[20] Mark Fonaryov and Michael Lindenbaum. On the minimal recognizable image patch. *CoRR*, abs/2010.05858, 2020. 3

[21] Wolf Garbe. Symspell: Spelling correction & fuzzy search - 1 million times faster through symmetric delete spelling correction algorithm, 2012. 5

[22] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 2

[23] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[24] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 1

[25] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*, pages 23885–23899. Curran Associates, Inc., 2021. 1

[26] Martin A. Giese and Tomaso Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192, 2003. 1, 2

[27] Melvyn A. Goodale and A.David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. 1, 2

[28] Liron Zipora Gruber, Shimon Ullman, and Ehud Ahissar. Oculo-retinal dynamics can explain the perception of minimal recognizable configurations. *Proceedings of the National Academy of Sciences*, 118, 2021. 3

[29] Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. 1

[30] Daniel Harari, Hanna Benoni, and Shimon Ullman. Object recognition at the level of minimal images develops for up to seconds of presentation time. *Journal of Vision*, 20(11): 266–266, 2020. 3

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016. 2

[32] John M. Henderson, Christine L. Larson, and David C. Zhu. Full scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: An fmri study. *Brain and Cognition*, 66(1):40–49, 2008. 2

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[34] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002. 2

[35] Boyue Caroline Hu, Lina Marsso, Krzysztof Czarnecki, Rick Salay, Huakun Shen, and Marsha Chechik. If a human can see it, so should your system: reliability requirements for machine vision components. In *Proceedings of the 44th International Conference on Software Engineering*, page 1145–1156, New York, NY, USA, 2022. Association for Computing Machinery. 3

[36] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959. 2

[37] Gokul Jacob, R. T. Pramod, Hemanth Katti, and S. P. Arun. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, 12(1):1872, 2021. 2

[38] Hyungjin Jang, Daniel McCormack, and Fangfang Tong. Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS Biology*, 19(12):e3001418, 2021. 2

[39] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2

[40] Saeed Reza Kheradpisheh, Mohammad Ganjtabesh, and Timothée Masquelier. Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition. *Neurocomputing*, 205:382–392, 2016. 3

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 1, 2

[42] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[43] Lisa K. Libby and Russell P. Eibach. How the self affects and reflects the content and subjective experience of autobiographical memory. In *The Self*, pages 75–91. Psychology Press, 2011. 2

[44] Grace W. Lindsay. Feature-based attention in convolutional neural networks. *CoRR*, abs/1511.06408, 2015. 1

[45] Margaret Livingstone and David Hubel. Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, 240(4853):740–749, 1988. 1

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[47] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 2

[48] Gaurav Malhotra, Marko Dujmović, and Jeffrey S. Bowers. Feature blindness: a challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, 18(5):e1009572, 2022. 2

[49] Timothée Masquelier, Thomas Serre, Simon Thorpe, and Tomaso Poggio. Learning simple and complex cells-like receptive fields from natural images: a plausibility proof. *Journal of Vision - J VISION*, 7:81–81, 2010. 3

[50] Heather K McIsaac and Eric Eich. Vantage point in episodic memory. *Psychonomic bulletin & review*, 9(1):146–150, 2002. 2

[51] David A. Mély and Thomas Serre. *Towards a Theory of Computation in the Visual Cortex*, pages 59–84. Springer Singapore, Singapore, 2017. 3

[52] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013. 5

[53] Romy Müller, Marcel Dürschmidt, Julian Ullrich, Carsten Knoll, Sascha Weber, and Steffen Seitz. Do humans and convolutional neural networks attend to similar areas during scene classification: Effects of task and image type. *Applied Sciences*, 14(6):2648, 2024. 3

[54] Nikolaas N. Oosterhof, Steven P. Tipper, and Paul E. Downing. Viewpoint (in)dependence of action representations: An mvpa study. *Journal of Cognitive Neuroscience*, 24(4):975–989, 2012. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2, 5

[56] Sadegh Rahmaniboldaji, Filip Rybansky, Quoc Vuong, Frank Guerin, and Andrew Gilbert. Dear: Depth-enhanced action recognition. *European Conference of Computer Vision 2024 WS*, 2024. 2

[57] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. 5

[58] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019 – 1025, 1999. Cited by: 2565. 2

[59] Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2 (9):661–670, 2001. 2

[60] Kevin Roche and Hugues Chainay. Visually guided grasping of common objects: effects of priming. *Visual Cognition*, 21 (8):1010–1032, 2013. 2

[61] Anna W. Roe. Modular complexity of area v2 in the macaque monkey. *Annual Review of Neuroscience*, 27:237–260, 2004. 2

[62] Anna W. Roe and Doris Y. Ts'o. Visual topography in primate v4: multiple maps of visual space. *Journal of Neuroscience*, 15(5):3689–3715, 1995. 2

[63] Filip Rybansky, Sadegh Rahmaniboldaji, Quoc Vuong, Andrew Gilbert, and Frank Guerin. Semantic consistency in identifying human actions. European Conference on Visual Perception, 2024. 5

[64] Thomas Serre. *Learning a dictionary of shape-components in visual cortex: comparison with neurons, humans and machines*. PhD thesis, Massachusetts Institute of Technology, USA, 2006. 3

[65] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429, 2007. 2

[66] Lior Shmuelof and Ehud Zohary. Dissociation between ventral and dorsal fmri activation during object and action recognition. *Neuron*, 47(3):457–470, 2005. 1, 2

[67] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 568–576, Cambridge, MA, USA, 2014. MIT Press. 1

[68] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, pages 1–14. Computational and Biological Learning Society, 2015. 2

[69] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 3

[70] Sanjana Srivastava, Guy Ben-Yosef, and Xavier Boix. Minimal images in deep neural networks: Fragile object recognition in natural images. *CoRR*, abs/1902.03227, 2019. 3

[71] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. 2

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[73] Felix A. Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9(Volume 9, 2023): 501–524, 2023. 2

[74] Daniel L. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. 1

[75] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. 1

[76] Huanjin Yao, Wenhao Wu, and Zhiheng Li. Side4video: Spatial-temporal side network for memory-efficient image-to-video transfer learning. *arXiv preprint arXiv:2311.15769*, 2023. 2, 4

[77] Boqiang Zhang, Kehan Li, Zhiqiang Hu Zesen Cheng, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2