

Human–AI Divergence in Ego-Centric Action Recognition under Spatial and Spatiotemporal Manipulations

Sadegh Rahmaniboldaji¹ Filip Rybansky² * Quoc C. Vuong² Anya C. Hurlbert² Frank Guerin¹ Andrew Gilbert¹

Received: date / Accepted: date

Abstract Humans consistently outperform state-of-the-art AI models in action recognition, particularly in challenging real-world conditions involving low resolution, occlusion, and visual clutter. Understanding the sources of this performance gap is essential for developing more robust and human-aligned models. In this paper, we present a large-scale human–AI comparative study of egocentric action recognition using **human-defined** Minimal Identifiable Recognition Crops (MIRCs), defined as the smallest spatial or spatiotemporal regions sufficient for reliable human recognition. We used our previously introduced, Epic-ReduAct, a systematically spatially reduced and temporally scrambled dataset derived from 36 EPIC-KITCHENS videos, spanning multiple spatial reduction levels and temporal conditions. **Recognition performance is evaluated using over 3,000 human participants, Side4Video, and Qwen 3.5 9B models.** Our analysis combines quantitative metrics, Average Reduction Rate and Recognition Gap, with qualitative analyses of spatial (high-, mid-, and

low-level visual features) and spatiotemporal factors, including a categorisation of actions into Low Temporal Actions (LTA) and High Temporal Actions (HTA). Results show that human performance exhibits sharp declines when transitioning from MIRCs to sub-MIRCs, reflecting a strong reliance on sparse, semantically critical cues such as hand–object interactions. In contrast, the model degrades more gradually and often relies on contextual and mid- to low-level features, sometimes even exhibiting increased confidence under spatial reduction. Temporally, humans remain robust to scrambling when key spatial cues are preserved, whereas the model often shows insensitivity to temporal disruption, revealing class-dependent temporal sensitivities. Overall, our findings demonstrate that strong benchmark performance can obscure fundamental human–AI misalignment and highlight limitations in current spatial and spatiotemporal representations, suggesting concrete directions for more efficient and human-aligned action recognition systems.

* Indicates these authors contributed equally.

Sadegh Rahmani
E-mail: s.rahmani@surrey.ac.uk

Filip Rybansky
E-mail: F.Rybansky2@newcastle.ac.uk

Quoc C. Vuong
E-mail: quoc.vuong@newcastle.ac.uk

Anya C. Hurlbert
E-mail: anya.hurlbert@newcastle.ac.uk

Frank Guerin
E-mail: f.guerin@surrey.ac.uk

Andrew Gilbert
E-mail: a.gilbert@surrey.ac.uk

¹ University of Surrey, Guildford, UK · ² Newcastle University, Newcastle upon Tyne, UK

Keywords computer vision · human vision · psychology · action recognition · ego-centric videos

1 Introduction

Ego-centric action recognition concerns the understanding of human actions from a first-person visual perspective and is fundamental to both human cognition and artificial intelligence (AI) models. It underpins a wide range of applications, including assistive technologies, augmented reality, human–computer interaction, and wearable AI. Effective ego-centric action understanding enables functionalities such as hands-free interaction, activity tracking, and robotic assistance.

In humans, action recognition is supported by two parallel visual processing streams: the ventral “what” and the dorsal “where/how” pathways (Giese and Poggio, 2003; Goodale and Milner, 1992; Shmuelof and Zohary, 2005). These pathways are hierarchically specialised to process complementary information, including form and motion. Similarly, early artificial intelligence models such as convolutional neural networks (CNNs) (Lecun et al., 1998; Krizhevsky et al., 2012) adopt hierarchical architectures, although they do not explicitly disentangle different types of visual information. More recent models, including two-stream networks (Simonyan and Zisserman, 2014), long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), and vision transformers (ViTs) (Dosovitskiy et al., 2021), aim to integrate spatial and temporal cues in a manner loosely inspired by human action recognition. However, despite these advances, such models remain vulnerable to occlusions and generally lack predictive processing mechanisms that enable humans to infer missing or ambiguous visual information (Fel et al., 2022).

Studies of hierarchically organised layered AI architectures have demonstrated that activations within these layers can predict neural responses in early (Güçlü and van Gerven, 2015), intermediate (Yamins et al., 2014), and late (Yamins and DiCarlo, 2016) areas of both monkey and human visual cortex. On this basis, some researchers argue that the combination of high task performance and alignment with behavioural and neural responses renders these models plausible approximations of human visual recognition (Lindsay, 2015; Bowers et al., 2023). Others, however, contend that their correspondence with human behaviour remains only slightly above chance (Geirhos et al., 2021, 2020b), raising questions about whether current models rely on the same information and strategies as humans.

Despite similarities in hierarchical architecture, the features extracted at each layer by human vision and AI models can differ substantially. For instance, in static visual recognition tasks, humans initially process boundaries and surfaces, which can result in a shape bias (Livingstone and Hubel, 1988), whereas AI models often exhibit a texture bias (Geirhos et al., 2019, 2020a; Malhotra et al., 2022). Additionally, AI systems frequently fail to encode the global configuration of visual elements (Baker et al., 2018), the three-dimensional and internal structure of objects, and cues related to occlusion and depth (Alcorn et al., 2019; Jacob et al., 2021; Dong et al., 2022; Bowers et al., 2023; Wichmann and Geirhos, 2023). Nonetheless, with specialised training datasets or architectural modifications, AI models can be trained to extract features that more closely resemble those utilised

by humans (Rahmaniboldaji et al., 2024; Geirhos et al., 2019; Jang et al., 2021).

Although both human vision and artificial models rely on hierarchical representations, humans uniquely integrate bottom-up sensory signals with top-down cognitive processes to achieve dynamic perception (Ahissar and Hochstein, 2004). Human perception is continually modulated by prior knowledge during inference (Bar, 2003; Hochstein and Ahissar, 2002), whereas deep learning models primarily depend on labelled data during training. Furthermore, human attention inherently prioritises salient regions based on contextual relevance (Torralba et al., 2006). This suggests that humans and current AI models may differ not only in the features they encode, but also in how they allocate and integrate spatial and temporal information when recognising actions.

Motivated by these findings and differences, we previously studied human performance in action recognition in (Rybansky et al., 2026) and conducted a range of quantitative analyses of spatial features in (Rahmaniboldaji et al., 2025). In this work, we build on and extend these contributions by further investigating the roles of spatial and spatiotemporal features in action recognition. We examine the differences between human observers and AI models in recognising actions in challenging ego-centric scenarios. Ego-centric videos provide detailed information about objects within an actor’s immediate ordinance (Borghi et al., 2012; Roche and Chainay, 2013) and their interactions via the hands (Campanella et al., 2011), while often offering limited contextual information (Henderson et al., 2008). This reduced context can affect both the visual features (Libby and Eibach, 2011; McIsaac and Eich, 2002) and cognitive strategies (Oosterhof et al., 2012; Rizzolatti et al., 2001) employed during action recognition, making ego-centric video an ideal domain for comparative analysis. We utilise videos from the standard ego-centric kitchen dataset EPIC-KITCHENS-100 (Damen et al., 2018) and systematically reduce available spatial and temporal information to extend the Epic-ReduAct dataset (Epic-Kitchens Reduced Action Videos) (Rahmaniboldaji et al., 2025; Rybansky et al., 2026), which is publicly released.

Although the analyses in this paper reveal that humans often exhibit larger recognition drops than AI models when critical information is removed, humans remain substantially more accurate overall on the original action-recognition task. As shown in Table 1, human observers achieved higher recognition accuracy than both S4V (Yao et al., 2023) and Qwen 3.5 9B (Team, 2025), particularly on challenging videos. Therefore, the motivation for studying human-AI alignment is not to

¹ Epic-ReduAct is available [here](#).

Table 1: Baseline action-recognition accuracy (%) on the original, unmanipulated videos. Human observers outperform both S4V and Qwen 3.5 overall, particularly on the challenging Hard subset, motivating our investigation of the cues and strategies underlying successful human recognition.

Set	Human	S4V	Qwen 3.5
Easy	100.0	88.89	61.11
Hard	100.0	55.56	27.78
Overall	100.0	72.22	44.44

reproduce human failure modes, but to understand the visual cues that support superior human recognition. By identifying the minimal spatial regions and their temporal ordering sufficient for reliable human recognition and comparing them with the information used by current models, we can investigate whether behaviourally relevant cues are under-utilised by existing systems and potentially use such cues to develop more robust, efficient, and interpretable action-recognition models.

To quantify human and AI performance, and inspired by prior work on static images (Ben-Yosef et al., 2018) and third-person actions (Ben-Yosef et al., 2020), we adopt the concept of Minimal Recognisable Configurations (MIRCs), defined as the smallest spatial crops of a video that remain identifiable by humans. Further spatial reduction yields smaller sub-quadrants, which, when unrecognisable, are referred to as sub-MIRCs, (Rahmaniboldaji et al., 2025; Rybansky et al., 2026). In parallel, to explore the spatiotemporal constraints of human and AI models, MIRCs are temporally scrambled. For simplicity and to ensure comparison to their spatial equivalents, we also refer to these as sub-MIRCs, regardless of their recognisability.

Using Epic-ReduAct and the human-defined MIRC framework, we address the central question of *which spatial and spatiotemporal cues are minimally sufficient for humans and state-of-the-art models to recognise ego-centric actions*. We quantify performance disparities between humans and AI models using two complementary metrics: the Average Reduction Rate, which we introduced (Rahmaniboldaji et al., 2025; Rybansky et al., 2026), and the Recognition Gap, adapted from (Ben-Yosef et al., 2018). Beyond these quantitative measures, we conduct extensive qualitative analysis to examine feature variations across reduction levels and to introduce a novel categorisation of motion patterns. Together, these analyses provide insight into the key spatial and spatiotemporal differences between human and model-based action recognition and highlight concrete directions for improving AI models.

The main contributions of this paper are, therefore:

1. Proposing a diagnostic framework and **benchmark for evaluating model behaviour at the spatiotemporal boundary of human egocentric action recognition**, combining hierarchical spatial reduction, temporal scrambling, Epic-ReduAct, and two quantitative metrics that support systematic human-AI comparison under degraded visual conditions.
2. An actionable, qualitative, strategy-level spatial analysis establishing design principles for efficient and human-aligned video models. We introduce a spatial-efficiency guideline grounded in high- and mid-level feature analysis, showing that humans rely on sparse, semantically critical active object configurations, whereas a state-of-the-art video model depends primarily on distributed contextual objects and mid-level statistics, resulting in distinct failure and recovery patterns.
3. A qualitative, strategy-level spatiotemporal analysis distinguishing Low vs. High Temporal Actions, together with recommendations for shifting model reliance from incidental context toward behaviourally critical objects and interactions.

2 Related Works

2.1 Human and Computer Vision

Human action recognition is supported by two parallel neural processing streams, commonly referred to as the ventral and dorsal pathways (Giese and Poggio, 2003; Goodale and Milner, 1992; Shmuelof and Zohary, 2005). These pathways originate in the primary visual cortex (V1) and project to higher-level regions, including inferotemporal (IT) and prefrontal cortex (PFC) (Serre et al., 2007). Early stages of visual processing extract low-level features such as edges, colour, and motion (Hubel and Wiesel, 1959; Roe, 2004; Roe and Ts'o, 1995), which are subsequently integrated into increasingly complex spatial and spatiotemporal representations. According to the reverse hierarchy theory (Ahissar and Hochstein, 2004), rapid bottom-up categorisation is followed by top-down refinement, enabling precise and flexible action recognition under variable visibility and partial information (Bar, 2003; Bar et al., 2006; Hochstein and Ahissar, 2002). Motivated by these biological principles, recent AI models that incorporate motion-aware or attention-based mechanisms have achieved state-of-the-art performance in video action classification (Ahmadian et al., 2023).

In contrast, AI-based action recognition predominantly relies on deep learning models trained on large-scale datasets. Early approaches employed convolutional

neural networks such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2015), and ResNet (He et al., 2016), which were originally designed for static image analysis and therefore operate primarily on spatial information. Although their layered architectures were loosely inspired by human vision, their effectiveness for action recognition was limited, in part because temporal dynamics had to be approximated by stacking frames or using simple temporal pooling. More recent advances, including Transformers (Vaswani et al., 2017), Vision Transformers (ViTs) (Dosovitskiy et al., 2021), and contrastive learning frameworks such as CLIP (Radford et al., 2021), have substantially improved visual representation learning. These developments have been extended to video-specific architectures that more explicitly model motion, such as Side4Video (Yao et al., 2023) and MOFO (Ahmadian et al., 2023), which leverage motion-based self-supervised learning, as well as FILS (Ahmadian et al., 2024), which incorporates linguistic semantics into video representations.

With the integration of multimodal data, recent large language models (LLMs) and video-language models (VLMs), including VideoLLaMA3 (Zhang et al., 2025) and VideoChatGPT (Maaz et al., 2024), enable richer and more holistic scene understanding by combining spatial appearance, temporal evolution, and semantic context. However, despite these advances, relatively few works have systematically probed how such models trade off spatial detail and temporal information under explicit reductions or scrambling. Comparative analyses of human and computer vision offer valuable insights into both improving AI models and understanding human perception. While humans demonstrate strong generalisation and contextual reasoning abilities, particularly under degraded or incomplete spatial and temporal information, AI systems typically achieve high specific task accuracy but often lack adaptability. Bridging these gaps through biologically inspired architectures and controlled studies of spatial and spatiotemporal robustness represents a promising direction toward more robust artificial vision systems.

2.2 Biologically inspired Vision Models

Early neurocomputational models were strongly influenced by the hierarchical organisation of the human visual pathways. A seminal example is the HMAX model (Riesenhuber and Poggio, 1999; Jhuang et al., 2007; Serre et al., 2007), which processes retinal input through a neurophysiologically grounded hierarchy of alternating simple and complex cell layers, employing linear and nonlinear MAX-pooling operations. Giese et al. (Giese and Poggio, 2003) extended this framework to action

recognition, such as walking and running, by introducing a two-pathway feedforward architecture inspired by the human ventral and dorsal streams, explicitly modelling form and motion pathways. More recent work has sought to enhance HMAX-style models while remaining within biological constraints. These efforts include incorporating layers corresponding to a more complete anatomy of the visual cortex (Serre, 2006), modifying the sequence of information flow and weights according to context (Kheradpisheh et al., 2016; Masquelier et al., 2010), or exploiting the temporal coordination of neural activity (Mély and Serre, 2017).

These models emphasise the importance of hierarchical, parallel processing of spatial structure and temporal dynamics, and they provide valuable insights into biological mechanisms of visual recognition. However, they generally fall short of the performance achieved by modern deep learning-based vision systems described earlier, and only a small subset of this literature quantitatively examines how such models respond when spatial information is reduced or temporal structure is disrupted. Our work complements these biologically inspired approaches by using a modern video model and focusing on controlled manipulations of spatial and spatiotemporal information in natural ego-centric actions.

2.3 Joint AI and Human Studies

Research explicitly integrating human-inspired principles into computer vision has examined the role of informative image regions in improving recognition, particularly under partial occlusion or reduced spatial detail (Fonaryov and Lindenbaum, 2020), as well as methods for aligning neural network receptive fields more closely with human visual processing (Fel et al., 2022). Several studies have argued for aligning AI model behaviour with human perception, such that image manipulations imperceptible to humans should similarly have minimal impact on model predictions (Hu et al., 2022).

Comparative analyses have consistently demonstrated discrepancies between human observers and AI models in terms of visual attention and recognition performance on image classification tasks (Carrasco et al., 2020; Srivastava et al., 2019; Gruber et al., 2021; Harari et al., 2020; Müller et al., 2024; Dyck and Gruber, 2020). Notably, Ben-Yosef et al. (Ben-Yosef et al., 2018) introduced the concept of Minimal Recognisable Configurations (MIRC)s and sub-MIRC)s for static images, demonstrating that human recognition accuracy drops sharply once critical spatial features are removed, whereas AI models exhibit a more gradual degradation.

Extending this framework to video, (Ben-Yosef et al., 2020) identified the minimal spatial and temporal config-

urations required for human action recognition in third-person videos from the UCF101 dataset (Soomro et al., 2012). Their findings showed a substantially steeper recognition decline from MIRC to spatial sub-MIRC for humans than for AI models, while no significant human-model differences were observed for temporally downsampled sub-MIRCs. This suggests that humans may be more sensitive than current models to the removal of critical spatial structure, but not necessarily to moderate temporal reduction in the third-person setting considered.

Building on this comparative paradigm, our work focuses on ego-centric video, a domain that has received limited attention in this context, and examines a broader and more diverse set of ego-centric activities. We systematically manipulate both spatial information (through quadrant and sub-quadrant cropping) and spatiotemporal information (through temporal scrambling of MIRCs) to characterise the minimal configurations that remain recognisable to humans and to a state-of-the-art video model. In doing so, we provide a detailed comparison of how humans and AI systems trade off spatial versus temporal cues when recognising actions from natural first-person videos.

3 Methodology

Our research pipeline, shown in Figure 1, summarises the methodology used to compare human and AI performance in ego-centric video action recognition. We first apply a classifier to preselect videos into Easy and Hard subsets, thereby constructing a representative and balanced dataset spanning multiple difficulty levels. To facilitate a controlled comparison of recognition behaviour, we then systematically reduce spatial information or scramble video frames to disrupt temporal structure. Both human participants and an AI model are subsequently evaluated on these spatially reduced or temporally scrambled clips, enabling direct comparison of recognition performance and quantification of disparities between human and AI action recognition.

3.1 Problem Definition and Epic-ReduAct Preparation

To support our investigation, we constructed a balanced subset of the EPIC-KITCHENS-100 dataset (Damen et al., 2018), comprising *Easy* and *Hard* videos that represent distinct levels of difficulty for action recognition in AI models. Each subset consists of 18 videos with an average duration of 2.35 seconds (SD = 1.11 seconds), enabling controlled comparisons of human and AI recognition performance across varying difficulty levels.

The Easy and Hard subsets of Epic-ReduAct were derived using a state-of-the-art action recognition model (Ahmadian et al., 2023) to estimate class prediction probabilities. Videos with top-1 confidence scores above 60% were selected as candidates for the Easy subset. In contrast, videos in which the ground-truth label did not appear within the top-5 predictions were designated as Hard candidates. A subsequent manual curation step removed ambiguous or visually redundant samples to ensure clear separation between difficulty levels. **The resulting Epic-ReduAct dataset comprises actions across 11 verb classes in the Easy subset and 12 in the Hard subset. Although the final dataset contains only 36 source videos, these videos were selected from 22 distinct EPIC-KITCHENS environments, corresponding to multiple real-world kitchen layouts, participants, object arrangements, and recording conditions. As EPIC-KITCHENS was collected in participants’ native kitchens, the selected videos capture variations in scene structure, cooking styles, viewpoints, and illumination conditions. The resulting dataset statistics are summarised in Table 2.**

Next, we conducted online experiments in which the spatial information of the 18 Easy and 18 Hard videos (36 in total) was systematically reduced across eight hierarchical levels to identify Minimal Recognisable Configurations (MIRCs). The reduction procedure is illustrated in Figure 2 using an example video with the ground-truth label *close*. At Level 0, each video was spatially cropped to a region that best preserved the full extent of the action. At Level 1, each video was cropped into corner-centred quadrants, yielding four child sub-videos per parent: Upper-Left, Bottom-Left, Upper-Right, and Bottom-Right. Levels 2 through 7 were generated by recursively applying the same corner-cropping procedure to each parent video at the preceding level (see Figure 2).

Video labels for MIRCs and sub-MIRCs (Ben-Yosef et al., 2018) were assigned based on human recognition performance. Following the spatial division of a parent video into four quadrants, any quadrant recognised by at least 50% of human participants was selected for further recursive reduction. If none of the child quadrants exceeded this recognition threshold, the parent video was designated as a **MIRC**, and its unrecognisable child quadrants were labelled as **sub-MIRCs**.

Due to the exponential growth in the number of child quadrants at deeper reduction levels, we applied an efficient quadrant selection via overlap to restrict testing to the most informative candidates. This pruning substantially reduced experimental complexity while preserving the ability to reliably identify MIRCs. The procedure was applied as follows:

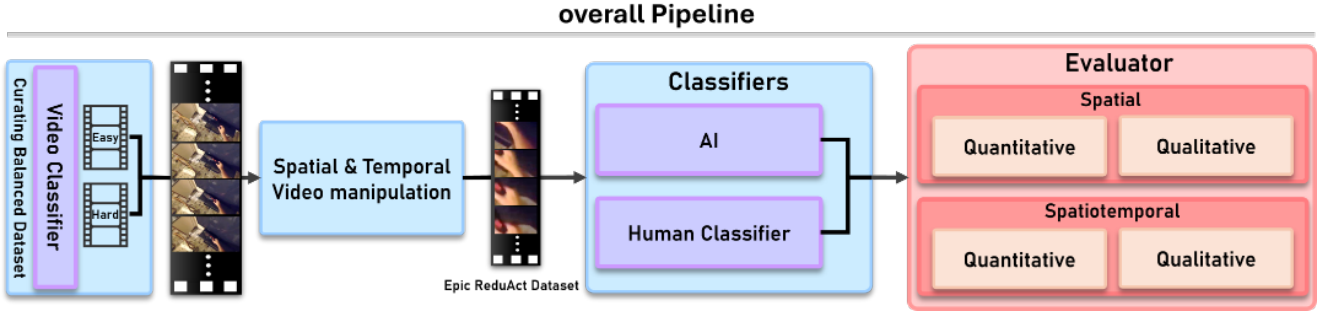


Fig. 1: Research pipeline highlighting different processes, including data preparation and classification by human and AI classifiers and the final analysis stage.

Category	Videos	Envs.	Samples	MIRCs	Spatial sub-MIRCs	Spatiotemporal quadrants	Verb Classes
Easy	18	15	4,503	273	1,092	273 (200)	close, cut, hang, open, pour, put, remove, take, turn-o , turn-on, wash
Hard	18	15	3,173	201	804	201 (145)	close, hang, insert, open, peel, pour, put, remove, serve, take, turn-o , wash
Overall	36	22 (similar envs)	7,676	474	1,896	474 (345)	–

Table 2: Summary of the Epic-ReduAct dataset. Values in parentheses indicate spatiotemporal unrecognisable quadrants.

1. After evaluating all quadrants at a given reduction level (see Section 3.2.2), child quadrants were generated only from parent quadrants that were recognised by at least 50% of participants.
2. Child quadrants whose spatial extent was fully contained within a parent quadrant that had been unrecognised at any previous level were assumed to be unrecognisable and were therefore excluded from further testing.
3. If child quadrants from a single level overlapped each other by at least 95%, we assumed that they had equal accuracy and tested only one quadrant from each such cluster.
4. Within each level’s remaining quadrants, we selected child quadrants that were (i) least likely to be recognised (sharing $\geq 65\%$ area with any unrecognised quadrant from previous levels) and (ii) most informative (showing the greatest cumulative overlap in surface area with other child quadrants at the same level). We continued this process until we reached the maximum number of quadrants per video that could be evaluated efficiently at the specified level.

Following the identification of spatial MIRCs through hierarchical cropping, we extended the reduction paradigm to the temporal domain to examine the role of temporal

structure in ego-centric action recognition. Specifically, we tested all MIRCs identified in the spatial reduction phase by selectively disrupting their temporal coherence while preserving spatial content. This allowed us to distinguish purely spatial MIRCs from *spatiotemporal* MIRCs, i.e., configurations for which both spatial and temporal information are critical for recognition.

To this end, each spatial MIRC video was temporally scrambled using a block-wise procedure designed to avoid excessive or unnatural discontinuities (Vuong and Tarr, 2004; Smekal et al., 2024). Each video was divided into five contiguous temporal blocks of approximately equal duration. The order of these blocks was then randomised under three constraints: (i) the first block could not remain in its original position, (ii) the last block was reassigned to a middle position, and (iii) no two blocks that were adjacent in the original video remained adjacent after scrambling. This manipulation preserved local motion statistics within blocks while disrupting the global temporal structure.

Recognition of temporally scrambled videos was evaluated using the same experimental protocol and decision criteria as in the spatial reduction phase. If scrambling caused recognition performance to fall below the 50% human recognition threshold, the configuration was con-

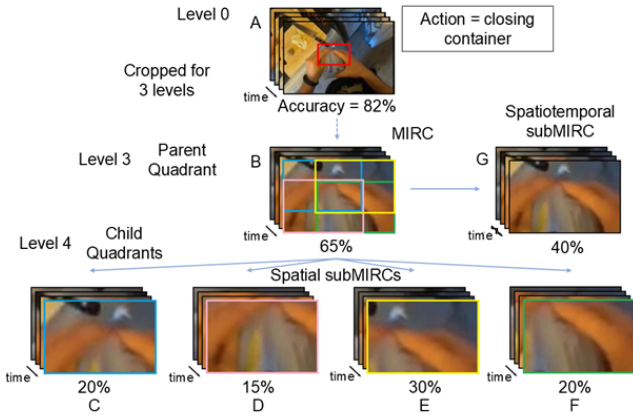


Fig. 2: Illustration of the reduction process. The ground-truth action label for the video is *close*. Panel A undergoes three successive levels of spatial reduction, resulting in Quadrant B, which is correctly recognised by 65% of participants. Quadrant B serves as the parent quadrant for four spatially reduced child quadrants, Upper-Left, Bottom-Left, Upper-Right, and Bottom-Right (C–F), each recognised by fewer than 50% of participants, thereby classifying them as sub-MIRCs and Quadrant B as the corresponding MIRC. In the spatiotemporal branch, the MIRC video is temporally scrambled, producing an additional child video (G), a spatiotemporal sub-MIRC, with a recognition rate of 40%.

considered unrecognisable. For simplicity, and to facilitate comparison with their spatial counterparts, we refer to these cases as spatiotemporal sub-MIRC, regardless of their recognisability status. Their corresponding unscrambled versions are classified as spatiotemporal MIRC (see Figure 2). Across all spatiotemporal MIRC configurations, 345 became unrecognisable, corresponding to 72.78% of the tested cases.

A single cohort of participants was recruited for this phase. Because the number of MIRCs varied across videos, we constructed multiple stimulus sets that were matched as closely as possible with respect to difficulty level (Easy/Hard), action category, and number of videos. Recognition testing was conducted in the same manner as in Phase 1, enabling direct comparison of spatial and spatiotemporal reduction effects.

All reductions, temporal scrambling, and pruning steps resulted in a total of 8,151 quadrant videos, including the Level 0 clips. The composition of the final Epic-ReduAct dataset is summarised in Table 2.

3.2 Classifiers

Using the spatially reduced and temporally scrambled videos, we evaluated action recognition using two classifiers: human observers, whose responses were also used to identify MIRCs and sub-MIRCs, and an AI model (Yao et al., 2023). Before describing these classifiers in detail, we first clarify how recognition accuracy is defined for each. For the AI model, accuracy corresponds to the confidence score assigned to the predicted verb relative to all other verb classes, expressed as a value between 0 and 1 (i.e., the softmax output). For human participants, accuracy is defined as the proportion of observers who correctly labelled the action shown in the video.

3.2.1 AI Classifiers

The AI classifier used in this study is based on the Side4Video (S4V) framework (Yao et al., 2023). This choice ensures that the balanced dataset of Easy and Hard videos is not biased by the MOFO-based selection process (Ahmadian et al., 2023), which relies on a fundamentally different, unsupervised training paradigm. Although MOFO and S4V achieve comparable performance, S4V offers substantially faster training, making it a more practical choice for integration into our pipeline.

As illustrated in the upper portion of Figure 3, S4V employs the OpenCLIP vision module (Radford et al., 2021) as a frozen, pre-trained video feature encoder, with a lightweight spatiotemporal side network attached. The trainable components operate in parallel with the frozen layers rather than being inserted sequentially, enabling efficient fine-tuning without back-propagation through the full model. This design leverages multi-level spatial features from the image encoder while significantly reducing memory usage (by up to 75% compared to prior adapter-based approaches) and allows large models such as ViT-E (4.4B parameters) to be effectively adapted for video understanding.

For our experiments, the model was trained on the EPIC-KITCHENS-100 training split comprising 67,179 videos (Damen et al., 2018), excluding the 36 videos reserved for human-AI comparison. Each video clip was uniformly subsampled to 8 frames at a spatial resolution of 224×224 . The model uses a ViT-B/16 backbone (Dosovitskiy et al., 2021) and was trained using the AdamW optimiser (Loshchilov and Hutter, 2019). The trained classifier was then applied to the spatially reduced and temporally scrambled videos in Epic-ReduAct to evaluate recognition performance against the ground-truth verb labels of the original videos, with results reported in Section 4.

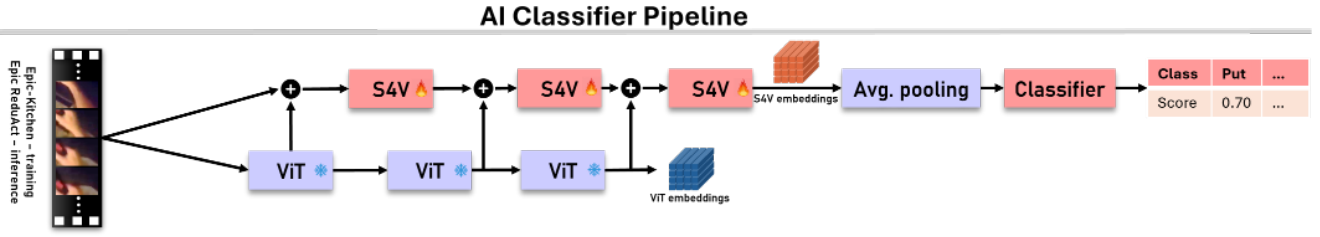


Fig. 3: The AI classifier (see Figure 1) is described in greater detail, consisting of a video feature encoder augmented by a lightweight spatiotemporal side network integrated with a frozen, pre-trained vision backbone.

During inference, the evaluation focuses on recognising ongoing action rather than identifying specific objects. Accordingly, verb recognition accuracy is computed by aggregating confidence scores across all correctly predicted verbs for each action label and comparing these aggregated scores with human recognition rates on the same spatially reduced and spatiotemporally scrambled clips.

3.2.2 Human Classifiers

The human classifier in this work is based on the experimental setup introduced in (Rybansky et al., 2026). Human recognition experiments were conducted using the Gorilla online platform (Anwyl-Irvine et al., 2020) and were distributed via Prolific (www.prolific.com). In total, data were collected from 4,360 participants (1,964 female, 2,329 male, and 67 identifying as other or not stated; mean age = 33.2 years, SD = 11.3 years), with 20 participants rating each cropped quadrant. Depending on the size of the assigned stimulus set, median completion times ranged from 9 to 21 minutes, with participant compensation ranging from £0.70 to £2.75, ensuring fair reimbursement. Ethical approval was obtained from the University Research Ethics Committee (Ref: 38465/2023).

The process of humans classifying a video is shown in Figure 4. Each experiment began with five practice trials. Participants were then presented with the balanced Easy/Hard dataset of 36 videos, and recognition accuracy was measured across trials. To control task difficulty and reduce learning effects, each participant viewed only one child quadrant per video, with trials presented in random order and interspersed with two catch trials. Practice and catch trials were drawn from additional Easy videos in the EPIC-KITCHENS-100 dataset (Damen et al., 2018). Seventeen participants were excluded and replaced due to poor performance on the catch trials.

Each trial began with a central fixation cross displayed for 500 ms, followed by a video presented on a

white background until a response was made. After 4000 ms, participants were prompted to type their response describing the observed action. Free-text responses were used to avoid bias introduced by predefined labels. Participants were instructed to identify a single action and the object being acted upon, using up to three words.

Human response analysis was based on an optimised Semantic Similarity measure, S_{sim} , relative to a human-GT label determined by (Rybansky et al., 2026). Responses were first cleaned by removing punctuation, articles and generic subjects (e.g. “man”, “person”), correcting misspellings algorithmically using SymSpell (Garbe, 2012), and manually rewording responses with incorrect word counts or slang.

The cosine similarity, CS , between each cleaned response and corresponding human-GT label was computed using the sentence-BERT (SBERT) model (*all-mpnet-base-v2*) (Reimers and Gurevych, 2019), and, in addition, the similarity between isolated verb terms (action), CS_A , and object terms, CS_O , was computed using Word2Vec embeddings (Mikolov et al., 2013). The final Semantic Similarity S_{sim} score was computed as:

$$S_{sim} = CS - (CS_O \cdot p)^2 + (CS_A \cdot b)^2,$$

where p and b denote penalty and bonus constants, respectively, which were optimised in advance on a separate validation set. A response was classified as correct if its S_{sim} score exceeded a predefined threshold calibrated to match manual labelling, and human recognition accuracy for each video was defined as the proportion of participants whose responses were classified as correct (Rybansky et al., 2026).

To validate this automated scoring procedure against human judgement, we conducted an independent study with 22 Prolific participants. They evaluated 576 unique pairs comprising a human-GT label and a free-text response sampled from the reduction experiments. For each of the 36 source videos, 16 responses were selected, balanced between responses classified by the automated scorer as correct and incorrect. Participants each rated a balanced set of 288 pairs, such that every pair received

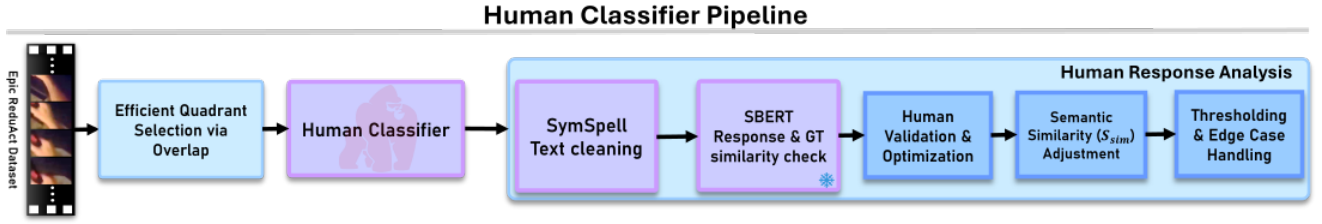


Fig. 4: The human classifier (see Figure 1) is presented in more detail, comprising two stages: classification and response cleaning.

Sensitivity	Speci city	Balanced Accuracy	Threshold
.923	.787	.855	.750

Table 3: Performance of the automated scorer relative to majority human consensus. The prespecified acceptance threshold was based on balanced accuracy.

11 independent judgements. They indicated whether the descriptions expressed the same action, prioritising the action over differences in the manipulated object.

For each pair, the majority human judgement defined the consensus label. The scorer was evaluated against this consensus using sensitivity, specificity, and balanced accuracy. To contextualise its performance against human variability, we also conducted a leave-one-rater-out analysis in which each participant was compared with the majority judgement of the remaining raters on the same items, producing a distribution of individual human balanced accuracies.

The scorer’s balanced accuracy against human consensus exceeded the pre-specified criterion of .750 (Table 3). Its balanced accuracy fell within the observed range of individual human raters and was close to their mean (.882), although below their median (.902; Figure 5). These results indicate that the automated scorer provides a reasonable approximation to majority human judgement and performs at a level observed among individual raters, supporting its use to classify free-text responses and derive human recognition rates in the reduction procedure.

4 Experiments and Results

In this section, we present a comprehensive set of experiments investigating the impact of feature manipulation in both spatial and spatiotemporal dimensions on human and AI action recognition. We use two quantitative metrics, the *Recognition Gap* and the *Average Reduction Rate*, to characterise how performance changes as in-

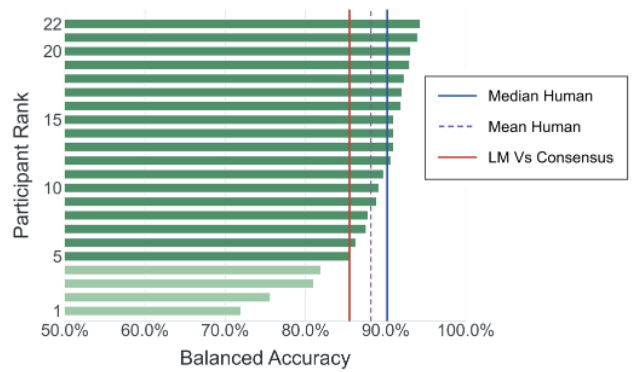


Fig. 5: Balanced accuracy of the automated scorer against majority human consensus relative to leave-one-rater-out balanced accuracy for individual human raters. Bars show individual participants; vertical lines show the automated scorer (solid red), human mean (dashed purple), and human median (solid blue).

formation is progressively removed. We further analyse which visual features persist or disappear across successive reduction levels, including temporal manipulations, to identify the features that induce transitions in model predictions. Together, these analyses enable a systematic study of how spatial and temporal cues affect both human observers and the AI classifier, and highlight features critical for recognition through complementary qualitative analyses. To improve the robustness of our findings, we additionally evaluated Qwen 3.5 9B as a second AI model. The corresponding experimental results are provided in the supplementary material (Section A.1 and A.2).

4.1 Quantitative Evaluation Metrics

We employ two complementary metrics to analyse the effects of spatial and spatiotemporal manipulations on action recognition performance: *Recognition Gap* and *Average Reduction Rate*. Both metrics are defined to enable direct comparison between human and AI recognition behaviour.

4.1.1 Recognition Gap Metric

The Recognition Gap quantifies how recognition performance changes between two consecutive levels of spatial reduction (Ben-Yosef et al., 2018; Rahmaniboldaji et al., 2025; Rybansky et al., 2026) or spatiotemporal scrambling (Rybansky et al., 2026), enabling a direct comparison between human observers and the AI classifier under matched baseline conditions. For human observers, the recognition gap is defined as the difference in average classification accuracy between parent MIRC and their corresponding sub-MIRCs. For the AI model, we define the recognition gap relative to human performance using a class-wise matched operating point. Specifically, for each action class, we first compute the average human recognition accuracy on MIRC videos, denoted by $X \in [0, 1]$. This value serves as a reference level for the task difficulty in that class. To compare the AI model at an equivalent operating point, we determine a class-specific confidence threshold tl such that the proportion of MIRC samples with predicted confidence exceeding tl equals X . In other words, tl is chosen to ensure that the AI model achieves the same effective recognition rate as humans at the MIRC level. **This calibration reduces dependence on absolute model confidence values and allows model behaviour to be evaluated relative to the human-defined MIRC boundary.** The threshold is illustrated as the green dotted line in Figure 8. We then calculate the average confidence gap between MIRCs equal or above the threshold and their corresponding sub-MIRCs. As an illustrative example, in Figure 8a, the average human accuracy for MIRCs is 59%. The corresponding AI threshold is defined as the confidence value above which 59% of the model’s MIRC predictions lie (e.g., a confidence of 0.15). The average gap for MIRCs above this threshold is 1.05%, indicating a slight improvement under spatial reduction for that class. This small change in the recognition gap arises because the metric first accounts for the AI model’s performance relative to human accuracy. Lower values indicate that the AI does not substantially outperform humans. Nevertheless, this metric allows us to identify improvements under reduction conditions, which in turn motivates further qualitative analysis.

By aligning AI evaluation to the human operating point on the parent videos, this metric isolates the effect of information reduction itself, allowing us to measure whether reduced spatial or temporally scrambled input **at the human recognition boundary** disproportionately harms or benefits the AI model relative to human observers.

4.1.2 Average Reduction Rate Metric

The Average Reduction Rate quantifies the reduction in recognition performance between two consecutive levels of spatial or spatiotemporal manipulation. For each parent–child pair, we compute the signed change in recognition accuracy a as

$$a = a_{\text{parent}} - a_{\text{child}},$$

where a_{parent} and a_{child} denote recognition accuracy in the range $[0, 1]$ for the parent and child, respectively.

To characterise the severity of recognition failures, we report the average reduction rate (ARR) over degrading transitions only, defined as the mean of positive accuracy drops, i.e. cases where recognition performance decreases under further reduction:

$$\text{ARR} = \frac{1}{N} \sum [a \mid a > 0].$$

This focuses the analysis on the magnitude and localisation of performance collapse as information is progressively removed, rather than conflating degradation with recovery.

4.2 Qualitative Evaluation Metrics

To qualitatively analyse the key features supporting recognition under spatial reduction, we extract two complementary feature sets: high-level object-based features and mid-level visual features. These features were specifically chosen to capture systematic changes in semantically and perceptually salient visual properties that are likely to affect recognition performance under progressive spatial reduction.

4.2.1 High-level object features

These features capture the proportion of semantically meaningful elements remaining in each quadrant relative to the original video, conceptually similar to occlusion-level metrics used in amodal segmentation studies (Qi et al., 2019; Zhu et al., 2017). As illustrated in Figure 6, segmented regions were grouped into four semantic categories: *Active Hand* (one or both hands directly involved in performing the action; Figure 6a), *Active Object* (up to two objects manipulated by the Active Hand and central to the action; Figure 6b), *Contextual Objects* (the combined area of all remaining segmented objects with potential contextual relevance, including the passive hand and forearms; Figure 6c), and *Background* (all image regions not covered by any segmented object).

We used the Segment Anything Model 2 (Ravi et al., 2024) (SAM v2.1_hiera_large) to produce initial segmentations. The model generated pixel-wise binary masks

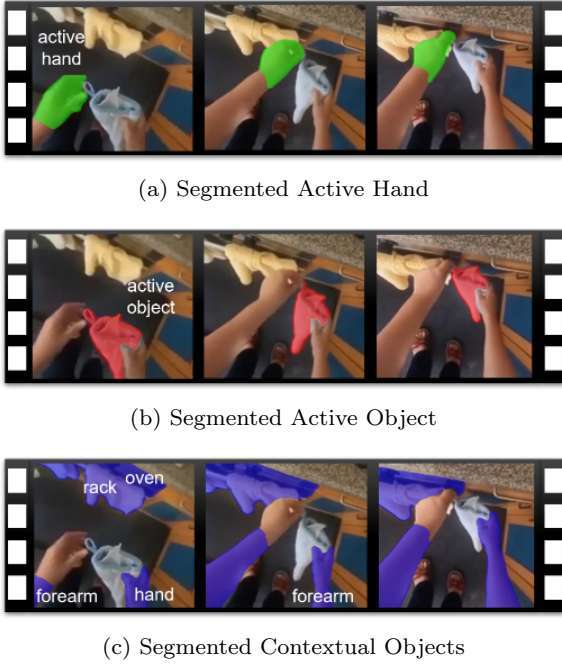


Fig. 6: Example segmentations illustrating the Active Hand (Figure 6a), Active Object (Figure 6b), and Contextual Objects (Figure 6c) in a video depicting the action labelled as “hang gloves”.

for each semantic category in each frame, assigning 1 to pixels belonging to that object and 0 to all others. Frame-wise masks were then temporally aggregated, and suboptimal masks were further refined manually using a custom Python-based annotation tool.

4.2.2 Mid-level visual features

Different configurations of mid-level features compound into high-level features and drive human attention in the initial stages of processing to potentially informative objects or regions. The features used here were obtained using the Graph-Based Visual Saliency (GBVS) algorithm, (Harel et al., 2006), which computes interpretable feature channels (DKL Colour, Intensity, Orientation, Colour, Flicker, Contrast, and Motion) aligned with the psychophysics literature and able to accurately predict human fixations when combined into a saliency map. Here we focused on the 7 mid-level conspicuity feature maps (Table 5). These maps provide normalised feature activation scores in the range $[0, 1]$ for each pixel, reflecting its activation relative to the surrounding region.

We next analysed the visual drivers of recognition failure and recovery shown in Figure 12 by tracking the visual components encoded in mid- and high-level feature maps across spatial reductions. We did this by quantifying the proportion of each object’s surface area

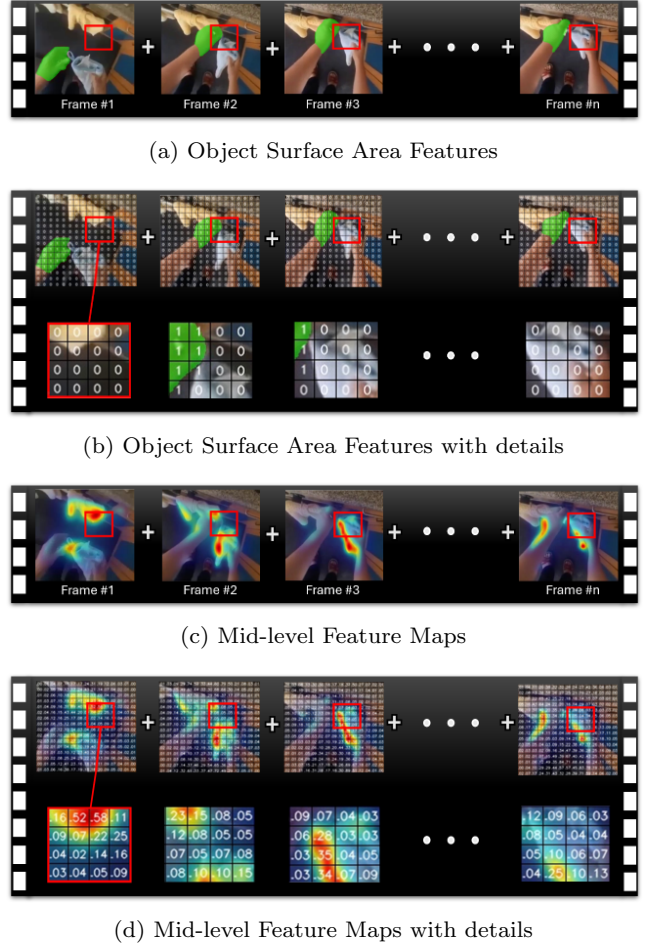


Fig. 7: Computation of Object Surface Area Features (Figure 7a) and Mid-level Feature Maps (Figure 7c)

and each GBVS feature activation score preserved in the quadrant after cropping, relative to the full video. As illustrated in Figure 7, we first summed pixel-wise values for every single frame inside the cropped quadrant, $M_{q,p}$, and the full video, $M_{f,p}$, then summed those across all frames of the quadrant, S_q , and the full video, S_f :

$$S_q = \sum_{q=1}^Q \sum_{p=1}^P M_{q,p}$$

$$S_f = \sum_{f=1}^F \sum_{p=1}^P M_{f,p}$$

Here, $M_{q,p}$ and $M_{f,p}$ denote either binary surface-area values (for object-based features) or continuous activation values in the range $[0, 1]$ (for mid-level features) at pixel p , within the cropped quadrant or full frame, respectively. For temporally scrambled quadrants, GBVS conspicuity maps used for $M_{q,p}$ were recomputed after frame shuffling, while $M_{f,p}$ remained based on the full unscrambled video.

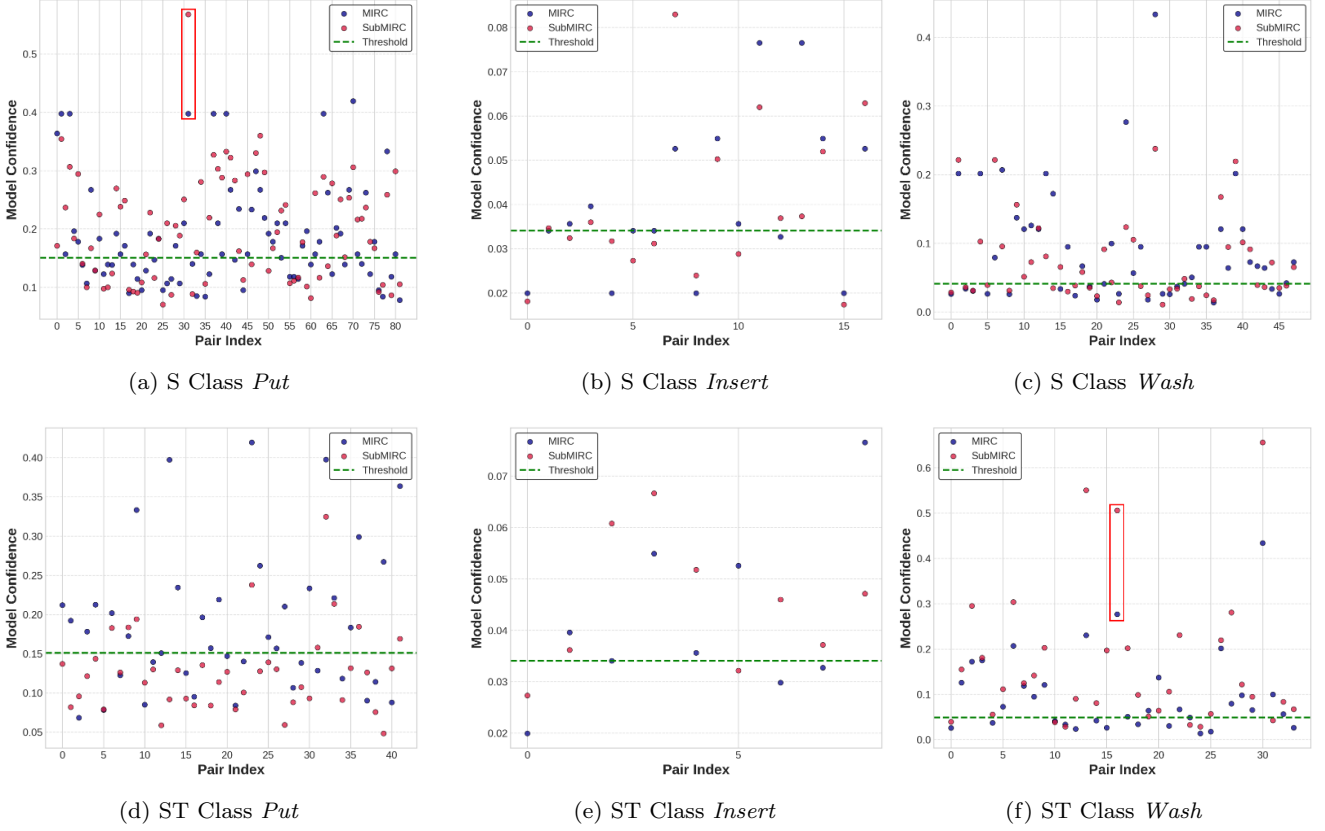


Fig. 8: Class-wise recognition gap for the AI model relative to a human-calibrated threshold. The recognition gap is computed for each MIRC–sub-MIRC pair using class-specific acceptance thresholds derived from human recognition performance. The top row shows results for spatial reductions, while the bottom row correspond to spatiotemporal scramblings, denoted with S and ST, respectively. In the spatial setting, (Figures 8a to 8c) present three representative classes and sample 31 of the *put* class is highlighted by the red box and analysed in detail in Section 4 (see Figure 9). For the spatiotemporal setting, (Figures 8d to 8f) present the same classes as in the spatial case and sample 16 of the *wash* class is highlighted by the red box for further analysis in Figure 17.

The final feature value for each quadrant was obtained as the ratio

$$p = \frac{S_q}{S_f},$$

representing the proportion of total object surface area or feature activation retained after spatial reduction or temporal scrambling.

4.3 Spatial Results

To analyse the effect of multi-level spatial reduction on both human and AI model performance, we evaluate results from human observers and the S4V model (Yao et al., 2023) using the two aforementioned metrics. This analysis reveals nuanced differences in behaviour and robustness between the two predictor types.

4.3.1 Quantitative Spatial Results

a) Recognition Gap. This metric helps us understand the boundary of human recognition (MIRC and sub-MIRC) and compare how the AI model behaves at that same boundary. We evaluate the recognition gap across all action classes in Epic-ReduAct. Representative examples, including varying numbers of pairs, are shown in Figure 8a-8c to facilitate a direct comparison between human performance and the AI model. These plots depict the relative positions of each MIRC/sub-MIRC pair with respect to the model’s confidence scores and the class-specific threshold line. A notable observation is that, in many cases, as shown in the highlighted sample in Figure 8a, sub-MIRC samples attain higher confidence scores than their corresponding MIRCs, which explains the negative recognition gap values shown later in Figure 10. This indicates that recognition performance can, in some cases, improve as spatial resolution is reduced.

		Classes													
Classifier	Metric	close	cut	hang	insert	open	peel	pour	put	remove	serve	take	turn-o	turn-on	wash
Human	RG	+42.92	+36.59	+40.52	+40.00	+38.07	+38.20	+32.21	+35.49	+31.36	+42.88	+39.03	+38.75	+31.11	+31.56
	LCI	+36.48	+32.20	+36.49	+31.97	+35.55	+31.31	+27.05	+31.99	+27.37	+36.41	+36.64	+33.31	+23.18	+26.53
	UCI	+49.35	+40.99	+44.56	+48.03	+40.59	+45.09	+37.36	+38.99	+35.36	+49.35	+41.41	+44.19	+39.05	+36.59
	N_v	3	1	2	1	5	1	2	5	2	1	5	3	1	4
	N_t	24	44	67	17	127	25	34	82	44	33	149	52	18	48
AI	RG	-2.30	+1.21	+0.35	+0.36	+5.40	+0.04	+0.07	-1.06	+5.71	+0.01	+15.12	+5.37	+19.44	+3.13
	LCI	-6.87	+0.20	-0.01	-0.54	+3.18	+0.02	-0.25	-3.45	+3.88	+0.01	+12.58	+3.03	+13.03	+0.67
	UCI	+2.27	+2.23	+0.70	+1.26	+7.63	+0.06	+0.38	+1.34	+7.55	+0.01	+17.66	+7.71	+25.84	+5.59
	N_v	3	1	2	1	5	1	2	5	2	1	5	3	1	4
	N_t	19	32	42	12	79	16	22	48	25	19	99	34	13	31

Table 4: **Spatial** Recognition Gaps (RG), lower and upper bounds of the 95% confidence interval (LCI and UCI), number of videos (N_v), and number of parent-child transitions between quadrants (N_t) for both classifiers (For AI model are only above threshold pairs), split by class label for spatial MIRC-sub-MIRC videos. RG and confidence-interval values are percentages. Positive RG values indicate a reduction in recognition performance under cropping, whereas negative values indicate an improvement. Confidence intervals were computed across individual transitions.

Feature	Definition
DKLcolour	Chromatic contrast in the Derrington-Krauskopf-Lennie colour space
Intensity	Luminance (brightness) variations
Orientation	Local edge orientation energy
Colour	Double-opponent Red-Green / Blue-Yellow contrasts
Flicker	Frame-to-frame luminance change (temporal contrast)
Contrast	Local brightness variability
Motion	Apparent motion energy between consecutive frames derived from optic flow

Table 5: GBVS feature channels and their definitions.

To further illustrate this effect qualitatively, we present an example video which shows how, in certain scenarios, spatial reduction can improve model performance rather than degrade it. Figure 9 shows an example video from the *put* class at three levels of reduction: the original video, the MIRC level (Level 2), and the sub-MIRC level (Level 3). At Level 2, the video still contains numerous irrelevant visual details. By contrast, at Level 3, removing such features enables the model to focus more effectively on the ongoing action. This focus improves action recognition, increasing the model’s confidence from 39% at the MIRC level to 56% at the sub-MIRC level. This example, similar to several other videos, corresponds to the red (sub-MIRC) and blue (MIRC) points highlighted in Figure 8a (pair 31).

Recognition gaps across all classes for both humans and the AI model are summarised in Table 4. Positive values indicate a reduction in recognition performance, whereas negative values indicate an improvement. The small negative values in the results indicate that MIRC and sub-MIRC do not define clear recognition boundaries for the AI model, unlike those defined by human observers. **Descriptive analysis shows that** human recog-

nition accuracy drops sharply across all classes. At the same time, the AI model generally shows improved or stable performance, as reflected by predominantly negative recognition gap values. For example, in the *close* class, human performance decreases by 42.92% ($\pm 6.44\%$), whereas the AI model improves by 2.30% ($\pm 4.57\%$). The most minor reduction in human recognition is observed for the *turn-on* class, which still shows a substantial drop of 31.11% ($\pm 7.93\%$), underscoring the strong effect of MIRC and sub-MIRC constraints on human perception compared to AI with 19.43% ($\pm 6.41\%$) drop in the same class which is the largest among other classes. **Class-level estimates should be interpreted cautiously because some verbs were represented by a small number of source videos.**

Figure 10 shows the frequency distribution of recognition gaps for human observers and the AI model. Our results exhibit distribution patterns that are consistent with prior image-based spatial reconstruction studies (Ben-Yosef et al., 2018). While AI models show modest improvements in recognition under spatial reduction, human accuracy consistently declines, with a markedly sharper drop than that of the AI models. Humans also

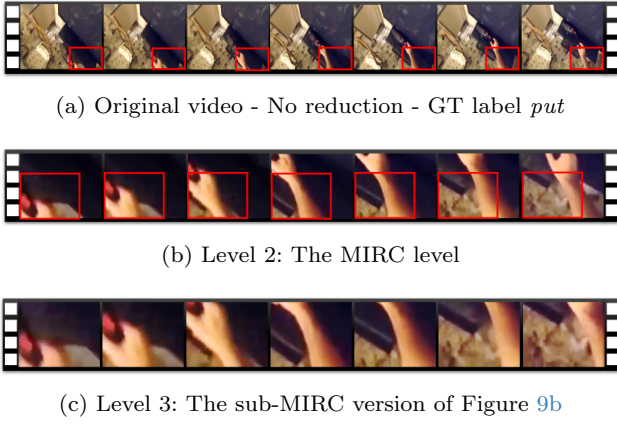


Fig. 9: Sample illustration of spatial feature loss across the original frame, MIRC, and sub-MIRC levels for a video from the *put* action class (corresponding to highlighted sample in Figure 8a). The red bounding boxes highlight the spatially reduced child video at the next reduction level. As spatial resolution decreases, background distractors are progressively cropped out, allowing the AI model to recognise the action correctly and thereby reducing the recognition gap between levels.

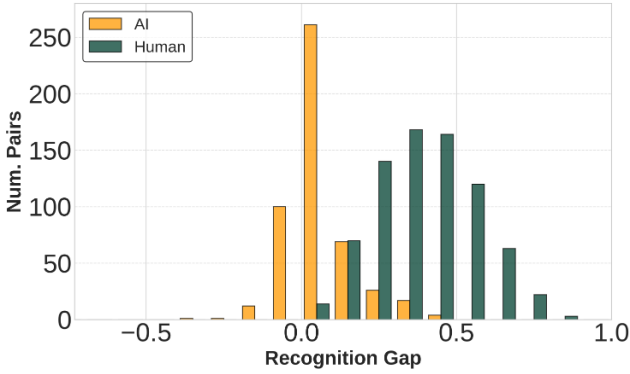
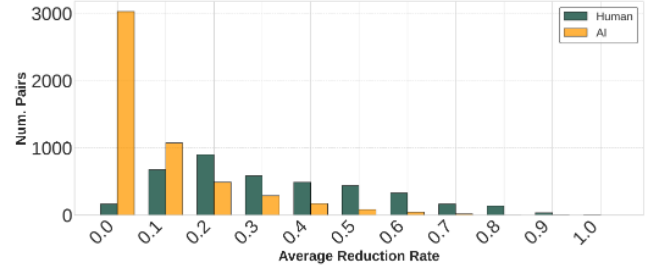
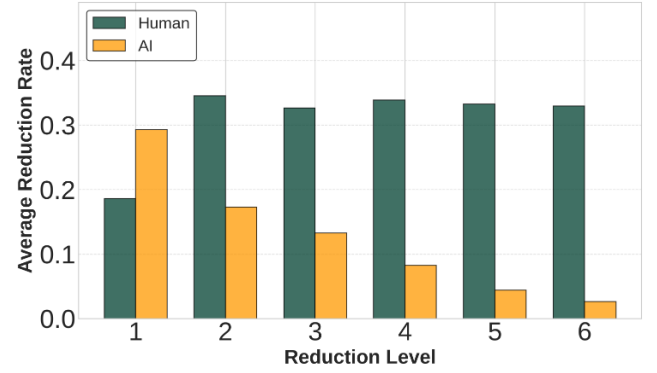


Fig. 10: Frequency distribution of recognition gaps between AI model and humans.

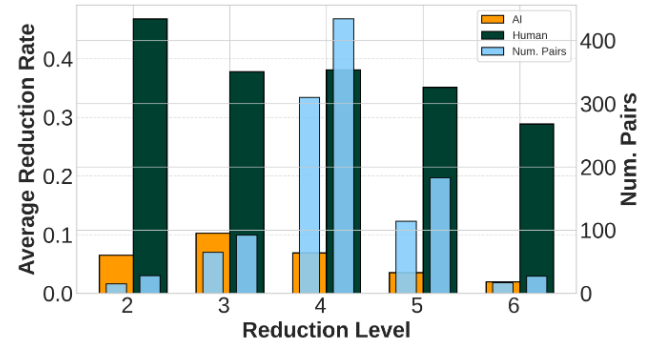
experience substantial losses in recognition confidence, whereas spatial reduction can improve the AI model’s action recognition, as reflected by negative recognition gaps (Table 4 and Figure 9). Moreover, the frequency distributions are broader for humans, indicating greater variability in performance degradation, while changes in the AI model are more gradual and concentrated. The 95% confidence intervals reinforce this contrast: for humans, the intervals remain entirely above zero across all classes, indicating a consistent reduction in recognition performance. For the AI model, several intervals overlap zero, suggesting no clear change in those classes, although classes such as *take* and *turn-on* show more pronounced reductions.



(a) Frequency distribution of average reduction rates across pairs of video quadrants.



(b) Average reduction rate (positive gaps) as a function of reduction level, restricted to MIRC-sub-MIRC pairs



(c) Average reduction rate as a function of reduction level

Fig. 11: Average reduction rate comparison between the AI model and humans. Figure 11a shows the distribution of parent-child pairs across all reduction levels, binned at intervals of 0.1 in reduction rate. Figure 11b shows the average reduction rate as a function of reduction level for all parent-child quadrant pairs. Figure 11c shows the same comparison, restricted to MIRC/sub-MIRC pairs, with the number of pairs visualised for each reduction level.

b) Average Reduction Rate. This metric enables comparison of model performance with results from psychophysical studies, highlighting levels at which recognition deteriorates significantly. Figure 11 presents human and AI classification performance. Figure 11a shows the frequency distribution of the average reduction rate for parent and child quadrants. Human and AI reduction rates exhibit different overall patterns as humans show a more pronounced decline in recognition, as reflected by a wider spread of higher reduction rates and more catastrophic failures across levels. To explain the larger number of pairs for the AI model compared to humans, this metric only counts pairs that show a reduction in recognition from parent video to child quadrant. This indicates that the AI experiences reductions in more pairs, though most are small, as reflected by the high frequencies of pairs in the 0.0–0.2 range. In contrast, humans exhibit larger reductions overall, with a greater number of samples showing substantial performance drops, as seen in both Figure 11a and Figure 11c.

Figure 11b shows the average reduction rate as a function of reduction level, highlighting the stages at which recognition declines most markedly. The figure indicates that the AI model experiences a gradual decrease in accuracy across levels, reflecting steady sensitivity to diminishing spatial information. In contrast, human recognition is robust to the first reduction level compared to the AI but deteriorates sharply from level 2 onwards as indicated by values exceeding 0.3.

Figure 11c reports the average reduction rate as a function of reduction level, restricted to MIRC-sub-MIRC pairs. In addition to performance trends, Figure 11c indicates the number of MIRC/sub-MIRC pairs available at each reduction level. This analysis highlights Levels 2 to 5 as the most critical stages for information loss for both humans and the AI model. However, while human recognition exhibits substantial and abrupt degradation at these levels, the AI model shows a more gradual decline, indicated by the generally lower average reduction rates less than 0.1. Overall, the trends observed in Figure 11c closely mirror those seen for all parent-child pairs in Figure 11b.

4.3.2 Qualitative Spatial Results

To complement the recognition-gap and reduction-rate results, we conducted an analysis of the features present in the quadrants to understand how and why action predictions change as the videos are progressively cropped. For each video, we construct a tree structure over its spatially cropped variants, as in Figure 2, but restricted to spatial reductions. Each node in the tree corresponds to a spatial quadrant of its parent video. We then tra-

verse each tree and record classification transitions by comparing the model’s predicted class (only the verb) at each parent-child pair, specifically identifying cases in which the verb prediction reverses between correct and incorrect responses. To ensure that these transitions reflect changes in action recognition rather than object detection errors, we analyse only pairs with a verb-annotation transition.

Figure 12 tracks how the presence of key spatial features changes across successive spatial reductions at points where action recognition changes. This feature drop analysis reveals two distinct mechanisms that govern recognition behaviour under spatial degradation: *feature loss*, where the removal of informative visual cues leads to recognition failure (Correct prediction → Incorrect prediction, shown in Figure 12a), and *noise reduction*, where suppressing distracting or irrelevant cues enables recognition recovery (Incorrect prediction → Correct prediction, shown in Figure 12b). Before analysing feature changes across transitions, we note a quantitative pattern not directly visible in the previous figures: Humans exhibit fewer prediction flips than the AI model, largely because the human experiment did not continue after failure cases. This was done to limit the number of tested quadrants and optimise the search for MIRCs. Specifically, the AI produced 866 transitions from correct to incorrect predictions and 257 from incorrect to correct, whereas humans exhibited 573 transitions from correct to incorrect. Although previous work has shown that human visual recognition can improve with reduced clutter (Levi, 2008), in a pilot experiment with the humans participants, they rarely recovered once a quadrant was not recognised, therefore we only consider AI in the recovery analysis in Figure 12b.

Failure via Feature Loss (Correct → Incorrect) For the AI model, our analysis first examined the magnitude of feature reduction when the AI transitions from a correct to an incorrect prediction. A first glance at Figure 12a, the values suggest that the loss of the *Active Object* (mean = 0.144) and *Active Hand* (mean = 0.129) are the primary drivers of failure, as these features experience the largest absolute reductions. However, this surface-level interpretation is misleading. A deeper statistical comparison reveals that the AI loses nearly identical amounts of the *Active Object* even when it succeeds ($\mu = 0.137$), with a negligible gap of only 0.007 between success and failure. Thus, the loss of high-level features (e.g., the main actor and action) are not the distinguishing features behind the AI failure.

Instead, a correlation analysis (Figure 13) reveals the true failure mechanism: *systemic collapse*. In fail-

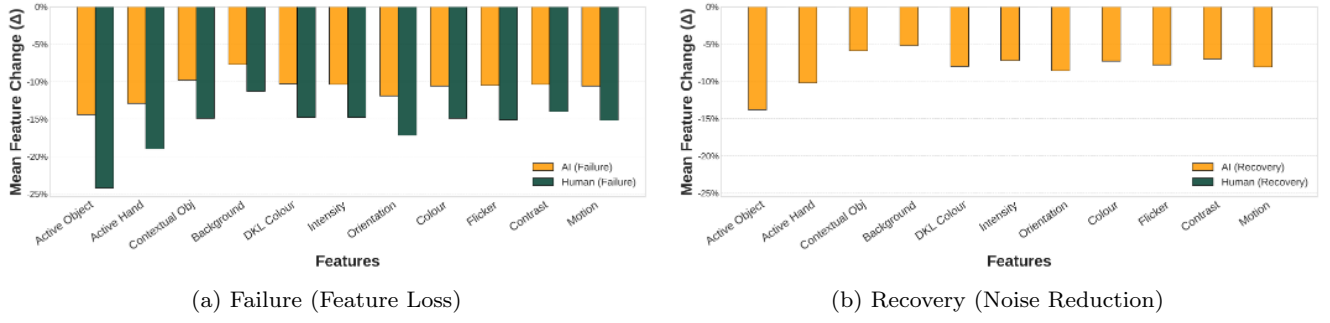


Fig. 12: Feature drop analysis. Quantifying the mid- to high-level features that lead to recognition failure and recovery. Figure 12a and Figure 12b show the mean feature change (Δ) in terms of feature visibility for Failure (Correct $\rightarrow$ Incorrect) and Recovery (Incorrect $\rightarrow$ Correct) transitions in human observers and the AI model. Note that human observers did not show any recovery transitions.

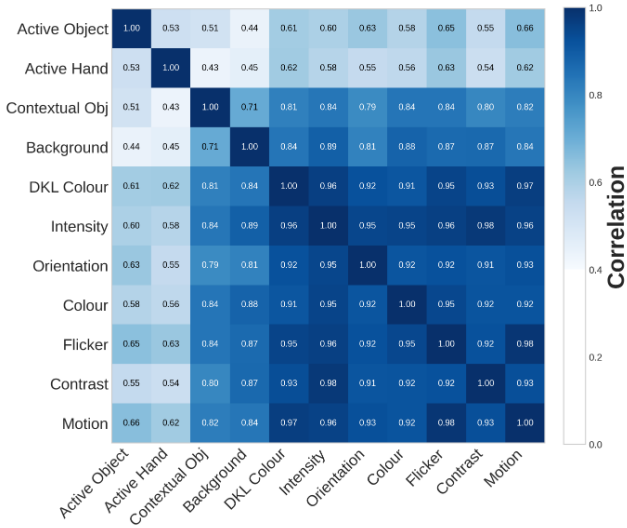


Fig. 13: Correlation matrix for AI failure cases, illustrating the strength of feature correlations across transitions in failure cases. (Correct $\rightarrow$ Incorrect).

ure cases, analysis of feature changes across transitions reveals extremely high correlations between the *Contextual Objects* feature and mid-level visual statistics, such as *Flicker* ($r = 0.84$), *Colour* ($r = 0.84$), and *Motion* ($r = 0.82$). Crucially, there is a moderate-to-strong correlation between the *Active Object* and *Contextual Objects* ($r = 0.51$), which is substantially higher than in success cases ($r = 0.38$). This indicates that AI failure is rarely caused by the loss of a single semantic cue. Rather, failure emerges when spatial reduction induces a synchronised collapse of the entire scene structure, wherein environmental context, lighting, motion, and the primary object degrade simultaneously.

This behaviour stands in contrast to human observers. For humans, failure is strictly *object-focused*.

The loss of the *Active Object* in human failure cases is substantial ($\Delta = 0.24$), nearly double the corresponding loss observed for the AI. This indicates that humans fail primarily when the semantic core of the action, the actor itself, is removed, and they remain comparatively resilient to the loss of background environment. This is consistent with human observers' tendency to use high-level features (Orban et al., 2021; Loucks and Baldwin, 2009). The AI, by contrast, exhibits brittle behaviour: it fails when the global statistical structure of the image collapses.

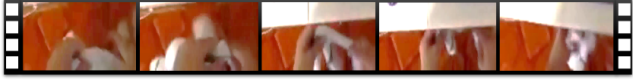
Figure 14 illustrates this distinction between human and AI behaviour for the representative *put* class.

As shown in Figure 14b with a lower-left crop of the original video, the AI flips to an incorrect prediction despite the *Active Object* and *Active Hand* being largely preserved by $\Delta = 0.04$ and $\Delta = 0.12$, respectively. Instead, the failure is driven entirely by the destruction of the scene: the *Contextual Objects* drop by 96%, while *Motion* feature activation is reduced by 57%. Although the primary actor remains visible, allowing human observers to correctly identify the action, the AI fails because the surrounding environmental anchors that support its prediction are removed.

We next consider a scenario in which another spatial quadrant is correctly classified to further isolate the contribution of the *Active Object* and *Active Hand*. Figure 14c illustrates this spatial quadrant from the same video at the same reduction level but obtained with an upper-corner crop. In this crop, the visibility of the *Active Object* and *Active Hand* is severely diminished, dropping to only 3% and 0.6%, respectively. In contrast, the *Contextual Objects* remain highly visible (79%), and the feature activation of colour, flicker, intensity, and motion, retain substantial presence (50%, 40%, 42%, and 31%, respectively). Despite the near-complete loss of the primary action cues, the model maintains a correct



(a) Original video (no reduction), verb correctly predicted as *Put*.



(b) Level 1 (Lower-left crop), verb incorrectly predicted as *Close*. *Contextual Objects* visibility drop by 96%.



(c) Level 1 (upper-right crop), verb correctly predicted as *Put*. 79% of *Contextual Objects* visibility and more than 30% activation of mid-level features.

Fig. 14: Example of a failure case under spatial reduction in a video in which the model produces an incorrect prediction after the first level of spatial reduction due to the complete occlusion of the *Contextual Objects*, while correctly recognising the action when the *Contextual Objects* and mid-level features remains visible.

prediction, highlighting that contextual and mid-level signals can occasionally compensate for missing object-hand information, albeit in a limited and case-specific manner.

Overall, these findings indicate that AI failures are predominantly driven by the degradation of environmental context such as contextual objects and also mid-level visual features, whereas human errors arise primarily from the loss of high-level semantic information related to the actor and main action.

Recovery (Incorrect $\rightarrow$ Correct) When analysing AI recovery cases, reductions in feature values again reveal a counterintuitive pattern. Successful crops do not necessarily maximise available information; instead, reduce it. In recovery scenarios, as shown in Figure 12b, the AI still suffers substantial loss of the *Active Object* ($\Delta = 0.132$). However, the critical differentiator lies in the preservation of the *Contextual Objects*, for which the reduction is minimal ($\Delta = 0.059$), compared to a substantially larger loss in failure cases ($\Delta = 0.097$). Although numerically small, this difference constitutes a performance *tipping point* for the model.

The correlation analysis characterises this recovery mechanism as *holistic preservation*. As shown in the correlation matrix in Figure 15, successful transitions exhibit a tight cluster of correlations across the transition between *Contextual Objects* and mid-level features such as *Orientation* ($r = 0.78$) and *Intensity* ($r = 0.77$).

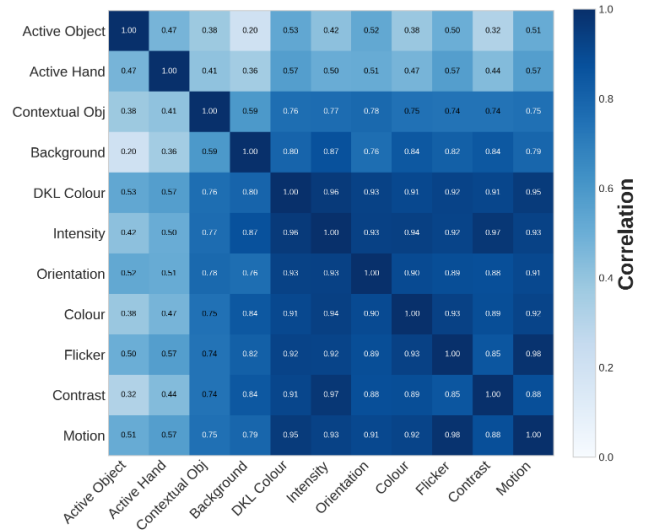


Fig. 15: Correlation matrix for AI recovery cases, illustrating the strength of feature correlations across transitions in recovery cases (Incorrect $\rightarrow$ Correct).

Crucially, the correlation between *Contextual Objects* and the *Active Object* weakens to $r = 0.38$. This relative independence supports a *pruning* hypothesis: the AI recovers not by retaining all available information, but by selectively preserving the environmental structure (context and orientation) while remaining largely insensitive to the partial loss of the *Active Object*.

While a direct comparison to a human ‘flip to correct’ is not possible, because we did not track the child quadrants of incorrect human classifications and thus did not explicitly measure recovery, the AI model nonetheless exhibits systematic feature amplification and noise suppression. If our assumption about the relative stability of human recognition failure is true, observers may rely primarily on their initial interpretation of the scene and require the restoration of the *Active Object* to correctly identify the action. By contrast, the AI can recover even when parts of the *Active Object* are missing, provided that the *Contextual Objects* are preserved. This indicates that the AI corrects its predictions by effectively “zooming in” on stable environmental cues rather than tracking the agent’s dynamics.

Similar to Figure 14, Figure 16 illustrates the differences between human and AI behaviour in a recovery scenario for the same representative *put* class. This transition serves as definitive evidence for *context pruning*. In this case, the AI transition from an incorrect (Figure 16b) to a correct (Figure 16c) action prediction despite a substantial reduction in the *Active Object* ($\Delta = 60\%$) and *Active Hand* ($\Delta = 51\%$). The decisive factor is the visibility of the *Contextual Object*,

which remains at 34% in Figure 16c. By removing potentially distracting foreground information, effectively pruning the object, and anchoring its reasoning in stable background cues (e.g., the sink and counter), the AI can refine and correct its prediction.

To further support this claim, consider Figure 9 again which is a representative example of the *put* class. In this example, the presence of the *Active Object* declines substantially from MIRC to sub-MIRC, falling to less than 2%. Although the *Active Hand* remains visible at $\approx 29\%$ in sub-MIRCs, this still represents a large decrease of $\approx 50\%$ compared to MIRCs. In contrast, *Contextual Objects*, as well as other mid-level features, largely preserve their presence, remaining at $\approx 22\%$ and within the 10%–15% range, respectively, with only minor reductions compared to the *Active Object* and *Active Hand*. This shift corresponds to an increase in model confidence from 39% in MIRCs to 56% in sub-MIRCs, and suggests that the AI model relies heavily on local textures and fine-grained cues which may be amplified through spatial cropping.

These recovery cases are consistent with a pruning interpretation: model confidence can improve when cropping removes foreground or locally salient visual information while preserving contextual scene structure. However, because cropping removes multiple spatially proximate features simultaneously, this correlational analysis cannot identify any single removed feature as the causal distractor. Confirming this would require targeted ablation or inpainting experiments in which the suspected distractor is removed while the background is kept intact.

Overall, recovery represents an act of *surgical pruning*. The model does not require a clearer view of the action itself to succeed; rather, it only needs an unobstructed view of the environment. Success is driven by the independence of background and foreground information, enabling the AI to reliably reach the correct prediction using stable environmental cues.

4.4 Spatiotemporal Results

This section examines how temporal context interacts with spatial reduction in action recognition for both human observers and the AI model. Overall, humans perform better than the AI model at recognising actions in temporally scrambled videos, correctly identifying 129 out of 474 samples, compared with 48 correct predictions by the AI model. However, this aggregate trend masks more nuanced findings, which we explore in detail in this section. To this end, we apply the same evaluation protocol used in the spatial analysis to the spatiotemporal setting, and report results for both human observers

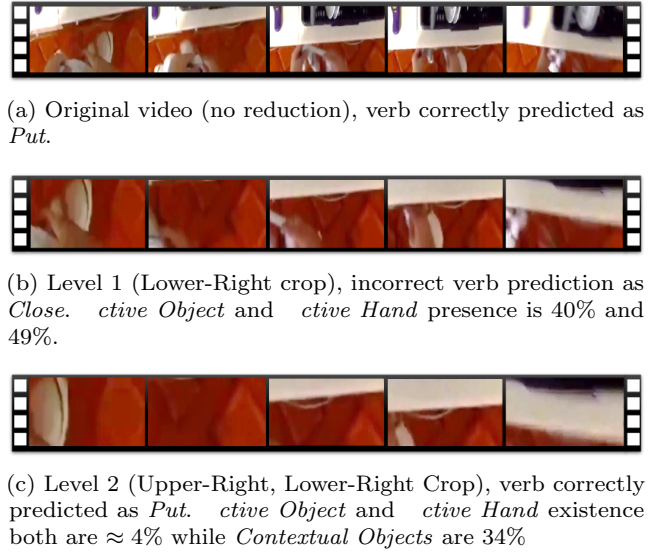


Fig. 16: Performance recovery via spatial reduction for a video with substantial *Active Hand* and *Active Object* suppression at Level 2 ($\approx 96\%$ compared to the original video). Preserving *Contextual Objects* and other mid-level features enables correct verb (action) recognition (*put*) after the initial misclassification.

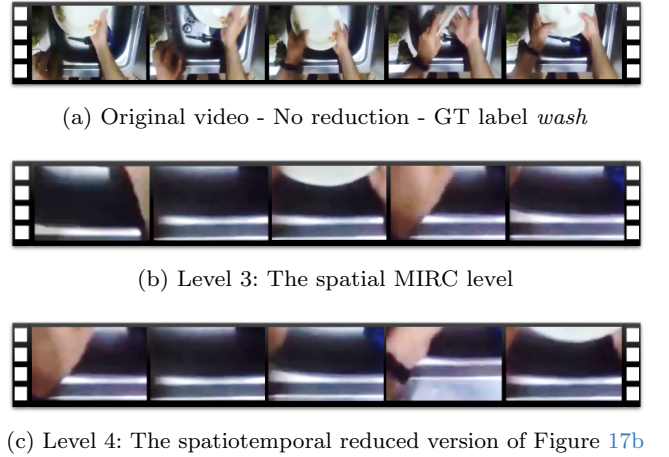


Fig. 17: An example of temporal scrambling applied to a spatial MIRC. Although the same frames are retained, their temporal ordering differs between Figure 16b and Figure 16c. As no spatial reduction is applied between the two levels, the observed differences arise solely from changes in temporal structure.

and the S4V model (Yao et al., 2023). Using the two previously introduced metrics, Recognition Gap and Average Reduction Rate, we analyse differences in behaviour and robustness between humans and the AI model with respect to temporal understanding.

4.4.1 Quantitative Spatiotemporal Results

a) Recognition Gap. We conduct quantitative experiments on spatiotemporal manipulations restricted to MIRC and sub-MIRC videos, as the most substantial changes occur at these stages; all preceding reduction levels are purely spatial and are therefore excluded from this analysis. The recognition gap is computed across all action classes in Epic-ReduAct. A subset of representative classes with varying sample counts is shown in Figure 8d–8f to facilitate comparison between human observers and the AI model. These figures depict the relative positioning of each MIRC/sub-MIRC pair with respect to the AI model’s confidence and the decision threshold. Consistent with the spatial-only results, the **spatiotemporal** sub-MIRC condition exhibits higher confidence than its corresponding **spatiotemporal** MIRC instance in many cases. This suggests that, similar to spatial reductions, temporal scrambling can sometimes preserve or improve action recognition performance. An example of this behaviour is illustrated by the red (sub-MIRC) and blue (MIRC) points in Figure 8f, highlighted by the red rectangle (pair 16), which experienced more than 20% improvement in confidence. To further examine this case, we visualise the corresponding videos in Figure 17. In Figure 17b, salient cues indicating the action *washing* (e.g., the presence of a sink and hand-object interactions) remain visible despite the shuffled frame order. As a result, no major loss of discriminative information is observed, indicating that temporal scrambling can preserve and, in some cases, enhance action recognition for the AI model. Similar patterns are also evident in several other videos across different classes.

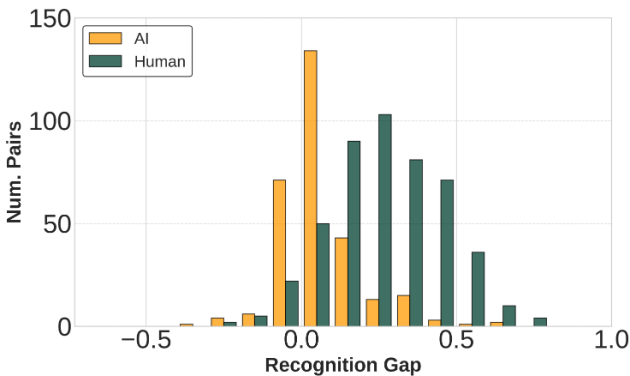


Fig. 18: Frequency distribution of the recognition gap between humans and the AI model under spatiotemporal reductions.

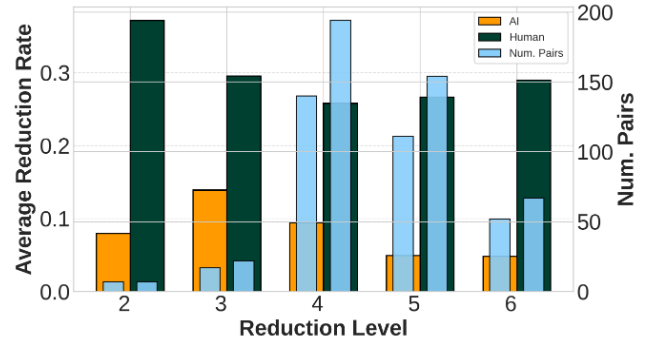


Fig. 19: Average reduction rate as a function of reduction level for MIRCs and spatiotemporal sub-MIRC pairs, along with the number of pairs in each level of reduction.

Notably, and less apparent in the per-class scatter plots, spatiotemporal manipulation can also improve performance for human observers in some instances. This reverse recognition gap effect is reflected in the aggregated results shown in Figure 18.

Consistent with the spatial analysis, AI models exhibit modest performance improvements under spatiotemporal manipulation (Figure 8); however, human observers also show limited gains following frame scrambling. Despite these improvements, most samples show a decline in recognition performance, with the drop substantially more pronounced for humans than for AI models, as indicated by the higher positive values.

The corresponding frequency distributions further emphasise this disparity: human performance shows a broader spread, reflecting more heterogeneous and occasionally abrupt changes in recognition, whereas AI models exhibit more gradual, concentrated variations. Taken together, these findings indicate that although recent advances in AI have narrowed certain aspects of the human-machine performance gap, substantial differences persist. Notably, based on Table 6, which indicates statistics of Figure 10 and Figure 18, the AI model’s mean remains close to zero and slightly negative (≈ -0.03), reflecting relative robustness and occasional improvements. A similar pattern holds under spatiotemporal reductions. In contrast, human observers exhibit substantially larger mean recognition gaps across all settings. Specifically, under spatiotemporal manipulations, human recognition degradation is more gradual than under purely spatial reductions, with the mean reduction being approximately 13% lower (37.3% for spatial compared with 24.7% for spatiotemporal) compared to the spatial condition. These contrasts between spatial reduction and temporal scrambling suggest that temporal scrambling is less disruptive to human recognition.

Manipulation Type	Classifier	Min	Max	Std.	Mean
Spatial	AI	0.425	+0.315	+0.112	0.037
	Human	0.050	+0.800	+0.159	+0.373
Spatiotemporal	AI	0.607	+0.320	+0.128	0.038
	Human	0.250	+0.750	+0.179	+0.247

Table 6: **Recognition gap statistics** (percentages) for human and AI classifiers under spatial and spatiotemporal reductions. Positive values indicate reduced recognition performance, while negative values indicate improvement.

		Classes													
Classifier	Metric	close	cut	hang	insert	open	peel	pour	put	remove	serve	take	turn-o	turn-on	wash
Human	RG	+32.35	+15.00	+35.63	+24.44	+23.56	+31.25	+19.05	+24.17	+20.00	+16.88	+26.69	+35.82	+15.00	+14.41
	LCI	+23.92	+9.28	+28.68	+18.23	+19.92	+20.49	+13.96	+18.87	+14.09	+10.66	+23.07	+29.64	+4.14	+8.51
	UCI	+40.79	+20.72	+42.57	+30.65	+27.20	+42.01	+24.14	+29.46	+25.91	+23.09	+30.31	+42.00	+25.86	+20.31
	N_v	3	1	2	1	5	1	2	5	2	1	5	3	1	4
	N_t	17	22	32	9	80	12	37	42	29	16	80	49	14	34
AI	RG	+3.59	+0.04	-1.53	-0.02	+14.45	+0.08	+0.82	+10.11	-2.51	-0.00	+15.52	+5.67	+18.77	-7.52
	LCI	-2.34	-0.41	-2.65	-1.78	+10.29	+0.06	+0.67	+7.27	-6.80	-0.01	+11.63	+2.77	+6.04	-11.77
	UCI	+9.51	+0.49	-0.41	+1.73	+18.60	+0.11	+0.97	+12.95	+1.77	-0.00	+19.41	+8.56	+31.50	-3.27
	N_v	3	1	2	1	5	1	2	5	2	1	5	3	1	4
	N_t	11	15	21	6	47	8	22	25	18	10	49	30	9	22

Table 7: **Spatiotemporal Recognition Gaps (RG)**, lower and upper bounds of the 95% confidence interval (LCI and UCI), number of videos (N_v), and number of parent-child transitions between quadrants (N_t) for both classifiers (AI model considered only above threshold pairs), split by class label for spatiotemporal MIRC-subMIRC videos. RG and confidence-interval values are percentages. Positive RG values indicate a reduction in recognition performance under scrambling, whereas negative values indicate an improvement. Confidence intervals were computed across individual transitions.

The range and variability statistics reveal further differences between human observers and the AI model. Human performance exhibits a substantially wider spread, with maximum recognition gaps reaching as high as +0.80 under spatial reductions and +0.75 under spatiotemporal reductions, indicating severe degradation in worst-case scenarios. In contrast, the AI model shows much narrower upper bounds, with maxima remaining below +0.32 across both settings. Similarly, the minimum values suggest that while both humans and the AI occasionally benefit from reductions, the AI model experiences larger improvements, particularly under spatiotemporal manipulations (minimum of -0.61), compared to humans (-0.25). This pattern is further reflected in the standard deviations, which are consistently higher for human observers, indicating greater variability and sensitivity to reduction conditions, whereas the AI model demonstrates more stable behaviour across both spatial and spatiotemporal settings.

Overall, the statistics in Table 6 reinforce two key findings: (i) spatial reductions are more damaging than spatiotemporal scrambling, especially for humans, and

(ii) humans exhibit larger degradation and greater variability than AI models.

In a more detailed, class-level **descriptive analysis**, the recognition gaps across all classes for spatiotemporal videos for human and AI models are shown in Table 7. The results indicate that spatiotemporal MIRC and sub-MIRC constraints continue to strongly affect human observers, while the AI model shows more stable or moderately improved performance. Human recognition accuracy drops noticeably across several action classes, with reductions reaching up to 35.82% ($\pm 6.18\%$) for the *turn-o* class or 15.00% ($\pm 10.86\%$) for the *turn-on* class, reflecting the variability in human detection under spatiotemporal distortions. The smallest decline among humans is observed in the *wash* class, with a modest 14.41% ($\pm 5.90\%$) drop, suggesting some robustness to specific actions. In contrast, the AI model demonstrates mostly minor changes in performance. Notably, the AI improves for the *wash* class (-7.52% $\pm 4.25\%$) and maintains relatively stable recognition for actions such as *cut*, *serve*, *peel*, *pour*, and *insert*. Even when performance declines, as in the *take* and *turn-on* classes (15.52% ($\pm 3.89\%$) and 14.45% ($\pm 4.16\%$), respectively), the re-

ductions are considerably smaller than those observed for human observers. **Class-level estimates should however be interpreted cautiously because some verbs were represented by a small number of source videos.**

The confidence intervals further illustrate the different response profiles of the two classifiers under spatiotemporal scrambling. For human observers, all class-wise confidence intervals remain positive, indicating a consistent reduction in recognition across action classes. The estimated gaps are largest for hang, turn-o , close, and peel, whereas smaller reductions are observed for wash, cut, turn-on, and serve. For the AI model, the pattern is less uniform. Some classes show confidence intervals clearly above zero, including open, put, take, turn-o , and turn-on, suggesting reduced model confidence after scrambling. In contrast, other classes show intervals below zero, such as hang and wash, indicating increased model confidence under temporal scrambling. Several remaining classes have intervals close to or overlapping zero, including close, cut, insert, remove, and serve, suggesting little or inconsistent effect of scrambling. Thus, unlike humans, the AI model does not show a uniformly negative effect of temporal scrambling across classes, but instead exhibits class-dependent increases, decreases, or near-zero changes in confidence.

Overall, the AI model appears less sensitive to spatiotemporal MIRC-subMIRC manipulations than human observers, highlighting differences in how temporal cues are integrated across actions **between classifiers**. Comparing the results for spatial (Table 4) and spatiotemporal (Table 7) MIRC-subMIRC videos reveals several key differences. For both the AI model and humans, recognition gaps are generally smaller in the spatiotemporal condition. **This suggests that the disruption of the global temporal continuity of spatial features is less destructive than the lack of spatial features themselves.**

b) Average Reduction Rate. Figure 19 presents the average reduction rate as a function of reduction level for MIRC/sub-MIRC pairs only, for both the AI model and humans.

In the analysis and comparison, as in the spatial domain, the superimposed bar charts (light blue) indicate that most pairs are concentrated at Levels 4 and 5. Additionally, the Average Reduction Rate trends across spatiotemporal reduction levels are strongly aligned with the spatial results, indicating that reduction levels have the same impact on changes in action understanding. However, the Average Reduction Rate values show that the average reduction is more severe for spatial manipulation for humans. Human reduction rates in the spatial domain consistently exceed those in the spatiotemporal

domain, peaking at over 0.45 at Level 2, Figure 11c, compared to a peak of approximately 0.37 in the spatiotemporal set, Figure 19.

An interesting finding for the AI model is that, although more samples in spatiotemporal experience improvements or maintain performance, reductions were slightly higher in spatiotemporal pairs, as shown by most of the AI-related (orange) bars. For instance, at Level 3, the AI average reduction rate reaches ≈ 0.14 in the spatiotemporal (Figure 19), noticeably higher than the ≈ 0.10 observed in the spatial domain (Figure 11c). A similar gap exists at Level 4, where spatiotemporal AI reduction is ≈ 0.10 compared to ≈ 0.07 for spatial.

4.4.2 Qualitative Spatiotemporal Analysis

We replicated the qualitative analysis used for spatial reductions in the spatiotemporal setting. However, this approach proved uninformative, as our high-level features were identical between spatial MIRC and scrambled sub-MIRC pairs. In contrast, mid-level features yielded negligible differences of 4.48×10^{-4} to 1.08×10^{-3} . **These negligible differences indicate that the paired videos were closely matched in cropped spatial content. Therefore, recognition differences are most directly attributable to the present block-wise temporal-order manipulation.**

Inspired by prior work (Bagad and Zisserman, 2025), which categorises actions based on their characteristics and inspired our hypothesis, we further examined the hypothesized class-level differences in classification outcomes following temporal scrambling. **The results revealed a clear class-dependent structure.** Based on this observation, we partitioned action classes into two verb categories: *Low Temporal Actions* (LTA) and *High Temporal Actions* (HTA), as summarised in Table 8. The analysis shows that HTA ($\approx 26\%$) yields substantially less improvement over scrambling for the AI model compared to LTA ($\approx 60\%$). A similar pattern is observed for human predictions; however, the magnitude of improvement is smaller. Figure 20a shows that the confidence delta for MIRC vs sub-MIRC recognition is higher for HTA than LTA classes, indicating that temporal scrambling not only degrades human accuracy for HTA classes but also leads to a marked reduction in subjective certainty. In contrast, LTA classes exhibit partial robustness to temporal perturbations.

By comparison, the AI model exhibits markedly different behaviour. For the model, the recognition gap is approximately 5% for HTA classes and 3% for LTA classes, suggesting that temporal scrambling has little impact on HTA recognition and may even improve performance on LTA classes. This pattern is mirrored

Verb Category	Verbs	Num. Pairs	Human Improved (Conf. < 0)		AI Improved (Conf. < 0)	
			Count	Percentage	Count	Percentage
HTA	opening, closing, turning, putting, taking, pouring, filling/serving, hanging, removing	406	21	5.17%	106	26.11%
LTA	washing, cutting, peeling	68	8	11.76%	41	60.29%

Table 8: Action categories grouped into High Temporal Actions (HTA) and Low Temporal Actions (LTA). The table reports the total number of pairs and the proportion of samples in which recognition improved after temporal scrambling for both human observers and the AI model.

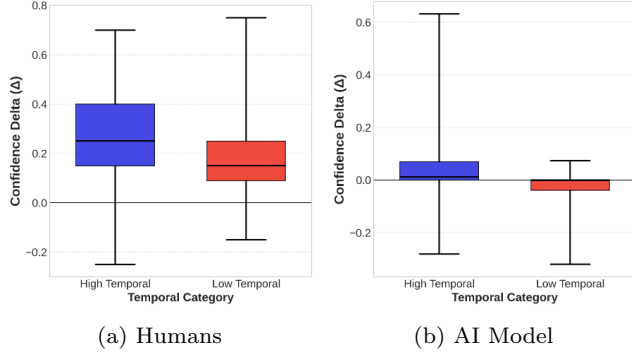


Fig. 20: Mean difference of Humans and AI Model prediction confidence between each MIRC and its spatiotemporal sub-MIRC.

in the confidence distributions shown in Figure 20b, where confidence deltas are tightly concentrated around zero, with minimal separation between HTA and LTA classes. Notably, LTA classes exhibit a slight negative shift in median confidence, indicating increased model confidence despite temporal disruption. A Welch’s t -test comparing the mean recognition gaps between LTA and HTA classes yields a significant difference for both human observers ($t(94.40) = 3.75$, $p = 3 \times 10^{-4}$, Cohen’s $d = 0.47$) and the AI model ($t(129.34) = 8.38$, $p = 10^{-4}$, Cohen’s $d = 0.81$). To ensure a more robust analysis and account for multiple clips originating from the same source video, we aggregate responses by unique video ID ($N = 36$) before comparing the groups, thereby treating each source video as the unit of analysis. Under this stricter evaluation, the difference remains statistically significant for the AI model with a large effect size ($t(14.04) = 4.12$, $p = 0.001$, Cohen’s $d = 1.21$), confirming that the inverted response to LTA scrambling is a robust model characteristic rather than an artefact of specific video instances. Conversely, the human response shows a trend toward greater temporal reliance for HTA classes, but this difference falls short of statistical significance ($t(8.61) = 2.15$, $p = 0.062$, Cohen’s $d = 0.82$) due to the reduced sample size. Ultimately, while humans exhibit a graded and interpretable response to temporal disruption across action categories,

the AI model responds differently to LTA and HTA classes, exhibiting distinct temporal-recognition characteristics that are largely invariant to temporal coherence. Additionally, further ablation studies on different frame sampling strategies using the S4V model are provided in the supplementary material (Section A.2.2).

5 Discussion and Conclusion

In this study we compared human and AI performance in ego-centric video action recognition, focusing on the contributions of mid- to high-level visual features to recognition. We used our new Epic-ReduAct dataset (Rahmaniboldaji et al., 2025; Rybansky et al., 2026), which contained a large number of video clips with systematically reduced spatial and temporal information. This dataset was derived from the EPIC-KITCHENS-100 dataset (Damen et al., 2018). Taken together, our findings suggest that the larger recognition drops observed in humans should not be interpreted as evidence of inferior robustness. Rather, humans achieve higher overall recognition performance while relying on a smaller set of semantically meaningful cues. This indicates that human recognition is often supported by compact and behaviourally relevant evidence, whereas current models frequently distribute evidence across broader contextual information. Understanding these differences provides a principled basis for designing future action-recognition systems that are both more effective and more interpretable.

5.1 Spatial cues and human–AI divergence

Our spatial analyses reveal a divergence between human observers and the evaluated AI model in how they respond to reductions around the human recognisability boundary. Humans show a sharp drop at the transition from MIRCs to sub-MIRCs, consistent with strong reliance on semantically critical spatial cues, particularly the configuration of the active hand–object interaction. By contrast, the AI model does not show the same

boundary-like behaviour at these human-defined transitions, instead showing more gradual degradation and, in some cases, improved confidence under further spatial reduction.

By contrast, the evaluated AI model shows a markedly more gradual degradation and, in many cases, even improved recognition under spatial reduction. This behaviour suggests that the model can benefit from removing potentially distracting information and relies primarily on distributed environmental context and mid-level visual statistics, rather than on the core semantic structure of the action. These findings demonstrate that strong performance on full-resolution benchmarks does not imply human-like spatial reasoning, and that recognition from minimal information is governed by qualitatively different strategies in humans and current models.

At the same time, our results suggest practical directions for improving the efficiency and performance of AI models. In particular, by incorporating either contextual objects or, human-relevant cues, such as active objects identified via human behaviour, as an auxiliary supervisory signal, we can guide models toward more meaningful representations. By explicitly encouraging models to attend to the same critical areas used by humans, it may be possible to shift model reliance away from incidental mid-level statistics and towards more semantic, human-aligned features, ultimately improving both robustness and interpretability.

5.2 Spatiotemporal cues and temporal robustness

Extending the MIRC framework to the temporal domain suggests that recognition of human-defined spatial MIRCs can be relatively robust to moderate block-wise temporal scrambling, particularly when critical spatial cues remain visible. Under this manipulation, humans show larger and more variable performance drops across action classes, whereas the AI model shows smaller changes and occasional confidence increases. This pattern is consistent with reliance on spatial or contextual cues and short-range motion statistics, but stronger temporal manipulations would be required to determine the extent to which the model uses coherent temporal order in full videos.

The present results therefore indicate that, within this human-defined MIRC setting, spatial reductions produced larger recognition losses than the block-wise temporal scrambling manipulation. They also suggest that action classes may differ in their dependence on temporal coherence. In particular, the distinction between Low Temporal Actions (LTA) and High Temporal Actions (HTA) provides an exploratory way to describe

how different actions respond to the present scrambling procedure. This distinction highlights a possible direction for future work: testing whether graded temporal manipulations, including stronger shuffling, reversal, and full-video scrambling, can identify when temporal order is necessary for human and model recognition.

5.3 Limitations

While our study provides a detailed characterisation of human–AI differences in ego-centric action recognition, several aspects also point to promising directions for future work. First, although the human behavioural study is large-scale, the high cost of experiments necessarily constrains the number of unique base videos. As a result, our findings are most directly applicable to the balanced subset of ego-centric actions examined here, and future studies could extend this framework to a broader range of activities. Second, our analyses focused on a single domain: ego-centric kitchen activities from EPIC-KITCHENS-100 (Damen et al., 2018), which are highly structured and object-centric. Although the hierarchical reduction procedure produced more than 8,000 clips, all stimuli were ultimately derived from 36 source videos. While these videos span 22 distinct EPIC-KITCHENS environments, they nevertheless represent only a subset of the variability present in real-world action-recognition settings. Extending the present paradigm to other ego-centric scenarios, or to third-person action recognition settings with richer global context, would provide a valuable test of the robustness of the observed human–AI differences across domains. Third, the spatial reduction tree was constructed using human recognition performance. MIRCs and sub-MIRCs therefore define a human recognisability boundary, and the human–AI comparisons should be interpreted as model behaviour at this human-defined boundary rather than as a symmetric estimate of the model’s own minimal recognition frontier. Future work could construct model-adaptive or exhaustively sampled reduction trees to quantify how human and model recognition boundaries overlap. Fourth, on the model side, we analysed a single state-of-the-art video architecture trained under a specific regime. Applying the same MIRC-based reductions and spatiotemporal manipulations to a broader family of architectures and training objectives, including multimodal or more explicitly predictive models, would help to clarify which of the observed discrepancies reflect general limitations of current video models and which are specific to particular design choices.

5.4 Future work

Building on these findings and limitations, several avenues for future work emerge. A first direction is to use Epic-ReduAct and the MIRC framework as a training signal rather than only as an evaluation tool. For example, AI models could be explicitly optimised to maintain performance across MIRC and sub-MIRC transitions, encouraging them to rely on the same semantically critical cues as human observers. A further promising line of work is to extend the proposed reduction and scrambling paradigm to a broader family of architectures and learning setups, including multimodal video language models and more explicitly predictive models. Systematically comparing how such models behave under controlled spatial and spatiotemporal reductions would help to disentangle architectural from data-driven effects and guide the development of architectures whose internal representations and failure modes more closely mirror human perception.

More broadly, we see MIRC-based spatial and spatiotemporal manipulations as a general toolbox for probing and shaping action recognition systems. Applying this toolbox across datasets and domains has the potential to yield models that are not only accurate on benchmarks, but also better aligned with human strategies for recognising actions in complex, ego-centric environments.

6 Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7 Data Availability Declaration

Data will be made available on reasonable request. Or adjust as appropriate.

Acknowledgements

This work was funded by the Leverhulme Trust Research Project Grant RPG-2023-079.

References

- Ahissar, M. and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, 8(10):457–464. [24](#), [25](#)
- Ahmadian, M., Guerin, F., and Gilbert, A. (2023). Mofo: Motion focused self-supervision for video understanding. *NeurIPS 2023 Workshop Self-Supervised Learning: Theory and Practice*. [25](#), [26](#), [27](#), [29](#)
- Ahmadian, M., Guerin, F., and Gilbert, A. (2024). Fils: Self-supervised video feature prediction in semantic language space. *British Machine Vision Conference (BMVC'24)*. [26](#)
- Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. (2019). 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. [24](#)
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52:388–407. [30](#)
- Bagad, P. N. and Zisserman, A. (2025). Chirality in action: Time-aware video representation learning by latent straightening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. [19](#), [20](#), [43](#)
- Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):e1006613. [24](#)
- Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15(4):600–609. [24](#), [25](#)
- Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., and Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, 103(2):449–454. [25](#)
- Ben-Yosef, G., Assif, L., and Ullman, S. (2018). Full interpretation of minimal images. *Cognition*, 171:65–84. [12](#), [25](#), [26](#), [27](#), [32](#), [35](#)
- Ben-Yosef, G., Kreiman, G., and Ullman, S. (2020). Minimal videos: Trade-off between spatial and temporal information in human and machine vision. *Cognition*, 201:104263. [25](#), [26](#)
- Borghi, A. M., Flumini, A., Natraj, N., and Wheaton, L. A. (2012). One hand, two objects: Emergence of a ordance in contexts. *Brain and Cognition*, 80(1):64–73. [24](#)
- Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., Puebla, G., Adolphi, F., Hummel, J. E., Heaton, R. F., and et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385. [24](#)
- Bulat, A., Perez Rua, J. M., Sudhakaran, S., Martinez, B., and Tzimiropoulos, G. (2021). Space-time mixing attention for video transformer. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19594–19607. Curran Associates, Inc. [18](#)
- Campanella, F., Sandini, G., and Morrone, M. C. (2011). Visual information gleaned by observing grasping movement in allocentric and egocentric perspectives. *Proceedings of the Royal Society B: Biological Sciences*, 278(1715):2142–2149. [24](#)
- Carrasco, J., Hogan, A., and Pérez, J. (2020). Laconic image classification: Human vs. machine performance. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20*, page 115–124, New York, NY, USA. Association for Computing Machinery. [26](#)
- Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. (2018). Scaling egocentric vision: The epic-kitchens dataset. *European Conference on Computer Vision (ECCV)*. [9](#), [24](#), [27](#), [29](#), [30](#), [44](#), [45](#)
- Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., and Zhu, J. (2022). Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. In *NeurIPS'22*. [24](#)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houtsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. [24](#), [26](#), [29](#)
- Dyck, L. and Gruber, W. (2020). Seeing eye-to-eye?: A comparison of object recognition performance in humans and deep convolutional neural networks under image manipulation. Workingpaper, University of Salzburg. 19 pages, 7 figures, 3 tables. [26](#)
- Fel, T., Felipe, I., Linsley, D., and Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432–9446. [24](#), [26](#)
- Fonaryov, M. and Lindenbaum, M. (2020). On the minimal recognizable image patch. *CoRR*, abs/2010.05858. [26](#)

- Garbe, W. (2012). Symspell: Spelling correction & fuzzy search - 1 million times faster through symmetric delete spelling correction algorithm. [30](#)
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020a). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673. [24](#)
- Geirhos, R., Meding, K., and Wichmann, F. A. (2020b). 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*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc. [24](#)
- Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. (2021). Partial success in closing the gap between human and machine vision. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23885–23899. Curran Associates, Inc. [24](#)
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*. [24](#)
- Giese, M. A. and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192. [24](#), [25](#), [26](#)
- Goodale, M. A. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25. [24](#), [25](#)
- Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (2017). The “something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850. [18](#)
- Gruber, L. Z., Ullman, S., and Ahissar, E. (2021). Oculoretinal dynamics can explain the perception of minimal recognizable configurations. *Proceedings of the National Academy of Sciences*, 118. [26](#)
- Güçlü, U. and van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014. [24](#)
- Harari, D., Benoni, H., and Ullman, S. (2020). Object recognition at the level of minimal images develops for up to seconds of presentation time. *Journal of Vision*, 20(11):266–266. [26](#)
- Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. In Schölkopf, B., Platt, J., and Ho, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press. [33](#)
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE. [26](#)
- Henderson, J. M., Larson, C. L., and Zhu, D. C. (2008). 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. [24](#)
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780. [24](#)
- Hochstein, S. and Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804. [24](#), [25](#)
- Hu, B. C., Marso, L., Czarnecki, K., Salay, R., Shen, H., and Chechik, M. (2022). 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*, ICSE '22, page 1145–1156, New York, NY, USA. Association for Computing Machinery. [26](#)
- Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591. [25](#)
- Jacob, G., Pramod, R. T., Katti, H., and Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, 12(1):1872. [24](#)
- Jang, H., McCormack, D., and Tong, F. (2021). Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS Biology*, 19(12):e3001418. [24](#)
- Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. [26](#)
- Kheradpisheh, S. R., Ganjtabesh, M., and Masquelier, T. (2016). Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition. *Neurocomputing*, 205:382–392. [26](#)
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. [24](#), [26](#)

- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. [24](#)
- Levi, D. M. (2008). Crowding—an essential bottleneck for object recognition: A mini-review. *Vision Research*, 48(5):635–654. [37](#)
- Libby, L. K. and Eibach, R. P. (2011). How the self affects and reflects the content and subjective experience of autobiographical memory. In *The Self*, pages 75–91. Psychology Press. [24](#)
- Lin, Z., Geng, S., Zhang, R., Gao, P., de Melo, G., Wang, X., Dai, J., Qiao, Y., and Li, H. (2022). Frozen clip models are efficient video learners. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, page 388–404, Berlin, Heidelberg. Springer-Verlag. [18](#)
- Lindsay, G. W. (2015). Feature-based attention in convolutional neural networks. *CoRR*, abs/1511.06408. [24](#)
- Livingstone, M. and Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, 240(4853):740–749. [24](#)
- Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*. [29](#)
- Loucks, J. and Baldwin, D. (2009). Sources of information for discriminating dynamic human actions. *Cognition*, 111(1):84–97. [38](#)
- Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. (2024). 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)*. [26](#)
- Malhotra, G., Dujmović, M., and Bowers, J. S. (2022). Feature blindness: a challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, 18(5):e1009572. [24](#)
- Masquelier, T., Serre, T., Thorpe, S., and Poggio, T. (2010). Learning simple and complex cells-like receptive fields from natural images: a plausibility proof. *Journal of Vision - J VISION*, 7:81–81. [26](#)
- McIsaac, H. K. and Eich, E. (2002). Vantage point in episodic memory. *Psychonomic bulletin & review*, 9(1):146–150. [24](#)
- Mély, D. A. and Serre, T. (2017). *Towards a Theory of Computation in the Visual Cortex*, pages 59–84. Springer Singapore, Singapore. [26](#)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. [30](#)
- Müller, R., Dürschmidt, M., Ullrich, J., Knoll, C., Weber, S., and Seitz, S. (2024). Do humans and convolutional neural networks attend to similar areas during scene classification: Effects of task and image type. *Applied Sciences*, 14(6):2648. [26](#)
- Oosterhof, N. N., Tipper, S. P., and Downing, P. E. (2012). Viewpoint (in)dependence of action representations: An mvpa study. *Journal of Cognitive Neuroscience*, 24(4):975–989. [24](#)
- Orban, G., Lanzilotto, M., and Bonini, L. (2021). From observed action identity to social affordances. *Trends in cognitive sciences*, 25(6):493–505. [38](#)
- Pan, J., Lin, Z., Zhu, X., Shao, J., and Li, H. (2022). St-adapter: parameter-efficient image-to-video transfer learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc. [18](#)
- Qi, L., Jiang, L., Liu, S., Shen, X., and Jia, J. (2019). Amodal instance segmentation with kits dataset. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3018. [32](#)
- Qing, Z., Zhang, S., Huang, Z., Zhang, Y., Gao, C., Zhao, D., and Sang, N. (2023). Disentangling spatial and temporal learning for efficient image-to-video transfer learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13934–13944. [18](#)
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020. [26](#), [29](#)
- Rahmaniboldaji, S., Rybansky, F., Vuong, Q., Guerin, F., and Gilbert, A. (2024). Dear: Depth-enhanced action recognition. *European Conference of Computer Vision 2024 WS*. [24](#)
- Rahmaniboldaji, S., Rybansky, F., Vuong, Q., Guerin, F., and Gilbert, A. (2025). Human vs. machine minds: Ego-centric action recognition compared. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2943–2953. [24](#), [25](#), [32](#), [44](#)
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*. [32](#)

- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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. Association for Computational Linguistics. 30
- Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019 – 1025. Cited by: 2565. 26
- Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9):661–670. 24
- Roche, K. and Chainay, H. (2013). Visually guided grasping of common objects: effects of priming. *Visual Cognition*, 21(8):1010–1032. 24
- Roe, A. W. (2004). Modular complexity of area v2 in the macaque monkey. *Annual Review of Neuroscience*, 27:237–260. 25
- Roe, A. W. and Ts'o, D. Y. (1995). Visual topography in primate v4: multiple maps of visual space. *Journal of Neuroscience*, 15(5):3689–3715. 25
- Rybansky, F., Rahmaniboldaji, S., Gilbert, A., Guerin, F., Hurlbert, A. C., and Vuong, Q. C. (2026). Seeing just enough: The contribution of hands, objects and visual features to egocentric action recognition. *bioRxiv*. 24, 25, 30, 32, 44
- Serre, T. (2006). *Learning a dictionary of shape-components in visual cortex: comparison with neurons, humans and machines*. PhD thesis, Massachusetts Institute of Technology, USA. 26
- Serre, T., Oliva, A., and Poggio, T. (2007). A feed-forward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429. 25, 26
- Shmuelof, L. and Zohary, E. (2005). Dissociation between ventral and dorsal fmri activation during object and action recognition. *Neuron*, 47(3):457–470. 24, 25
- Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14*, page 568–576, Cambridge, MA, USA. MIT Press. 24
- Simonyan, K. and Zisserman, A. (2015). 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. 26
- Smekal, V., Poyo Solanas, M., Fraats, E. I., and de Gelder, B. (2024). Differential contributions of body form, motion, and temporal information to subjective action understanding in naturalistic stimuli. *Frontiers in Integrative Neuroscience*, 18:1302960. 28
- Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402. 27
- Srivastava, S., Ben-Yosef, G., and Boix, X. (2019). Minimal images in deep neural networks: Fragile object recognition in natural images. *CoRR*, abs/1902.03227. 26
- Team, Q. (2025). Qwen3.5-9b. <https://huggingface.co/Qwen/Qwen3.5-9B>. Alibaba Cloud, Apache 2.0 License. 9, 14, 18, 24, 52, 53
- Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766. 24
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. 26
- Vuong, Q. C. and Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research*, 44(14):1717–1730. 28
- Wichmann, F. A. and Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9(Volume 9, 2023):501–524. 24
- Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365. 24
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624. 24
- Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., and Li, M. (2023). AIM: Adapting image models for efficient video action recognition. In *The Eleventh International Conference on Learning Representations*. 18
- Yao, H., Wu, W., and Li, Z. (2023). Side4video: Spatial-temporal side network for memory-efficient image-to-video transfer learning. *arXiv preprint arXiv:2311.15769*. 9, 20, 24, 26, 29, 34, 40, 52, 53

- Zhang, B., Li, K., Zesen Cheng, Z. H., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., and Zhao, D. (2025). Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*. [26](#)
- Zhu, Y., Tian, Y., Mexatas, D., and Dollár, P. (2017). Semantic amodal segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4542–4550. [32](#)

A Supplementary Material

This supplementary material provides additional experimental results supporting the findings reported in the main paper. We include quantitative and qualitative analyses using the Qwen 3.5 9B vision-language model, as well as ablation studies on temporal sampling strategies.

A.1 Additional Experiments on Quantitative Spatial and Spatiotemporal Analysis

Here, we include the quantitative results of Qwen 3.5 (Team, 2025), a recently developed vision-language model, to examine whether the observed human–AI behavioural divergence generalises beyond Side4Video (S4V) (Yao et al., 2023). Unlike S4V, which performs closed-set action recognition, Qwen 3.5 generates free-form responses that are evaluated using the same semantic alignment procedure as human annotations, providing an alternative evaluation protocol for comparison.

A.1.1 Recognition Gap

Figures 21 and 22 presents the Recognition Gap analysis for Qwen 3.5. Similar to S4V (Figure 10, Figure 18), most recognition changes are concentrated around small values, indicating that the model exhibits gradual changes in recognition across successive reductions. Compared with S4V, Qwen 3.5 shows slightly closer agreement with human behaviour, likely because both human responses and Qwen predictions are evaluated using the same prediction-alignment procedure. This is particularly evident in the small number of videos exhibiting large Recognition Gap values.

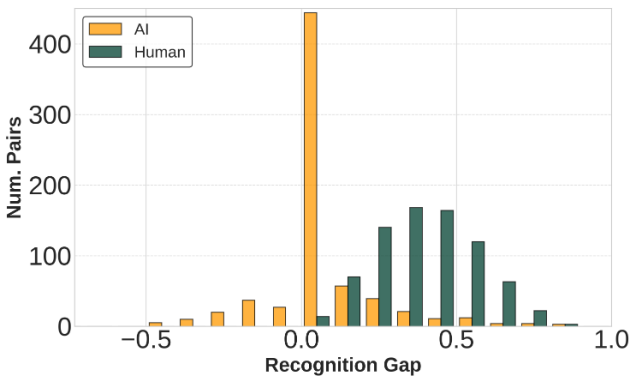


Fig. 21: Frequency distribution of the recognition gap between humans and the Qwen 3.5 (Team, 2025) under spatial reductions. Similar trend as S4V (Yao et al., 2023) but with wider range of positive and negative improvements.

A.1.2 Reduction Rate

On Spatial reduction rate, the Qwen 3.5 (Figures 23a to 23c) exhibits trends that closely match those observed for S4V (Figures 11a to 11c), with consistently small recognition drops across successive reductions. Despite their substantially different architectures and evaluation protocols, both models display remarkably similar reduction behaviours. In contrast, both remain clearly separated from the human trend, which

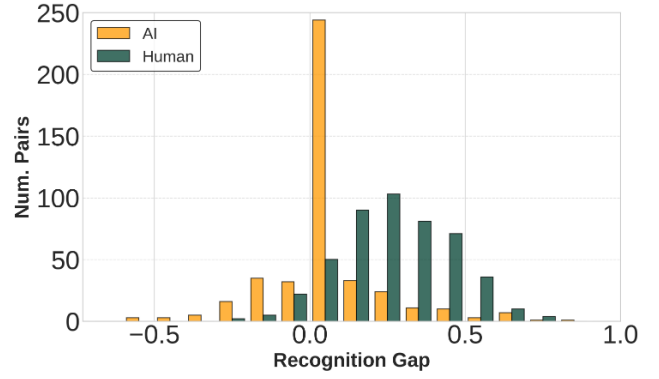


Fig. 22: Frequency distribution of the recognition gap between humans and the Qwen 3.5 (Team, 2025) under spatiotemporal reductions. Similar trend as S4V (Yao et al., 2023) but with wider range of positive and negative improvements.

exhibits substantially larger recognition drops. Among the two models, Qwen 3.5 shows slightly larger drops (approximately 0.2 per reduction level) than S4V (approximately less than 0.1), making its behaviour marginally closer to that of human participants due to similar response alignment to human while still following the same overall AI trend.

A similar pattern is observed for the spatiotemporal analysis. Qwen 3.5 (Figure 24) exhibits a reduction-rate trend that closely matches that of S4V (Figure 19), indicating that the observed behavioural patterns are consistent across substantially different model architectures and evaluation protocols.

A.2 Additional Experiments on Qualitative Spatial and Spatiotemporal Analysis

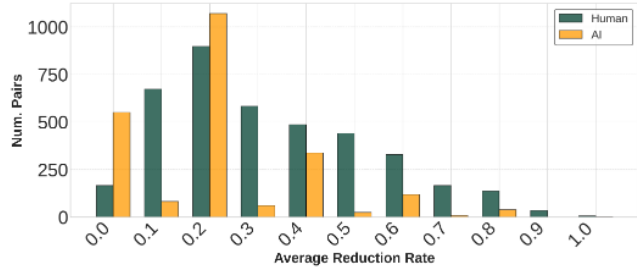
Here, we present qualitative results for Qwen 3.5 (Team, 2025), whose free-form responses are evaluated using the same semantic alignment procedure as human annotations, to examine whether its behavioural patterns relative to humans are consistent with those observed for S4V (Yao et al., 2023).

A.2.1 Spatial Analysis

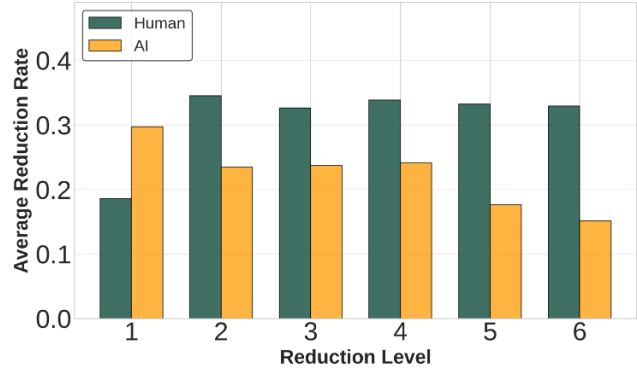
As shown in Figure 25a, the feature drop and recovery patterns closely match those observed for S4V (Figure 12a). Figure 25b presents the same analysis using Qwen 3.5’s raw free-form responses without semantic alignment to human annotations. The overall trends remain unchanged, indicating that the alignment procedure does not alter the qualitative behaviour. Instead, it primarily affects the absolute values, with the unaligned evaluation producing lower failure rates and higher recovery rates due to its less stringent matching criterion.

A.2.2 Spatiotemporal Analysis

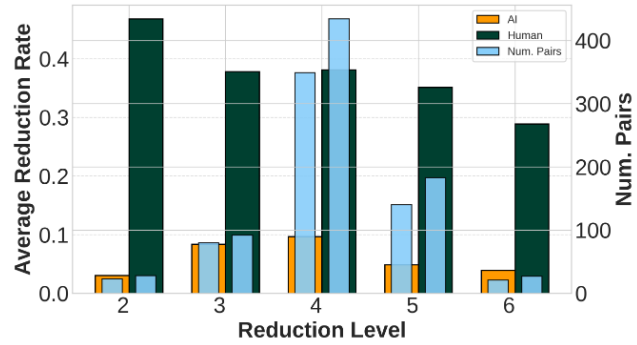
The comparative analysis of the AI’s confidence gaps across the three experimental setups using S4V model (Yao et al., 2023) strongly supports the findings of the initial temporal study. In both the 8-frame scrambled and 8-frame reverse evaluations, Low Temporal Actions (LTA) remained almost unchanged or even demonstrated improved model confidence, reinforcing



(a) Frequency distribution of average reduction rates across pairs of video quadrants.



(b) Average reduction rate (positive gaps) as a function of reduction level, restricted to MIRC-sub-MIRC pairs



(c) Average reduction rate as a function of reduction level

Fig. 23: Average reduction rate comparison between the Qwen 3.5 model (Team, 2025) and humans. Figure 11a shows a trend similar to that observed for the S4V model (Yao et al., 2023).

the hypothesis that cyclic, texture-based actions are highly resilient to temporal disruption. Conversely, the 1-frame inference approach (central frame) introduces a structural limitation to the study of temporal degradation. Distilling a video down to a single static image is no longer a mere reduction of temporal information; it constitutes the complete removal of temporal information (i.e. image changes over frame). Without the dimension of time, the model is stripped of its ability to perceive motion or causal sequence, forcing it to evaluate all actions as identical spatial classification tasks and entirely collapsing the categorical distinction between High and Low Temporal actions. Additionally, Table 9 reports the average

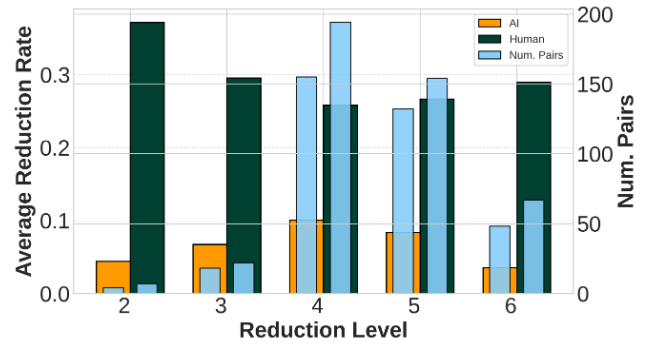
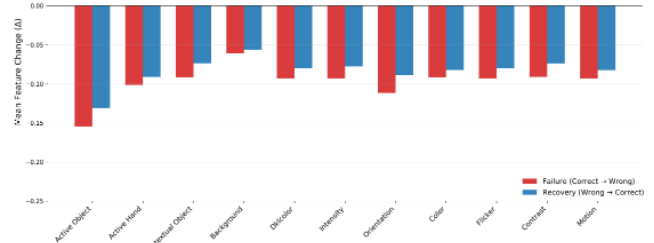
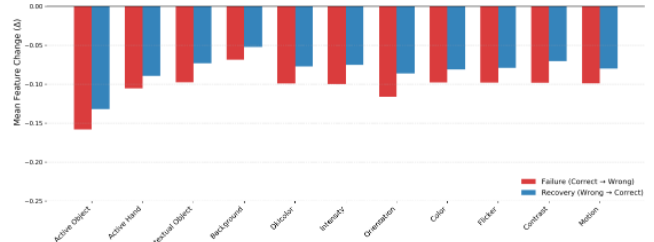


Fig. 24: Average reduction rate as a function of reduction level for MIRCs and spatiotemporal sub-MIRC pairs, along with the number of pairs in each level of reduction for Qwen 3.5 (Team, 2025) which is similar to S4V (Yao et al., 2023).



(a) Qwen 3.5 (Free range responses with response alignment)



(b) Qwen 3.5 (Free range responses)

Fig. 25: Feature drop and recovery analysis of Qwen 3.5 (Team, 2025), quantifying the mid- to high-level feature changes associated with recognition failure and recovery. Figure 25a shows the mean change in feature visibility () for Failure (Correct → Incorrect) and Recovery (Incorrect → Correct) transitions in Qwen 3.5 using the human prediction alignment approach discussed in Section 3.2.2, exhibiting a trend similar to that of the S4V model (Yao et al., 2023). Figure 25b presents the same analysis without the human prediction alignment approach.

performance across action categories under different temporal settings, reducing the effect of category imbalance. Across all settings, LTAs consistently exhibit either improvement or, at worst, lower performance degradation compare to HTAs.

Category	8-Frame Scrambled	8-Frame Reverse	1-Frame (No Temporal)
High Temporal ($n = 406$)	+0.052	+0.048	+0.052
Low Temporal ($n = 68$)	0.032	0.006	+0.033

Table 9: Sample-weighted average AI confidence gap per temporal category, across the three spatiotemporal ablations.

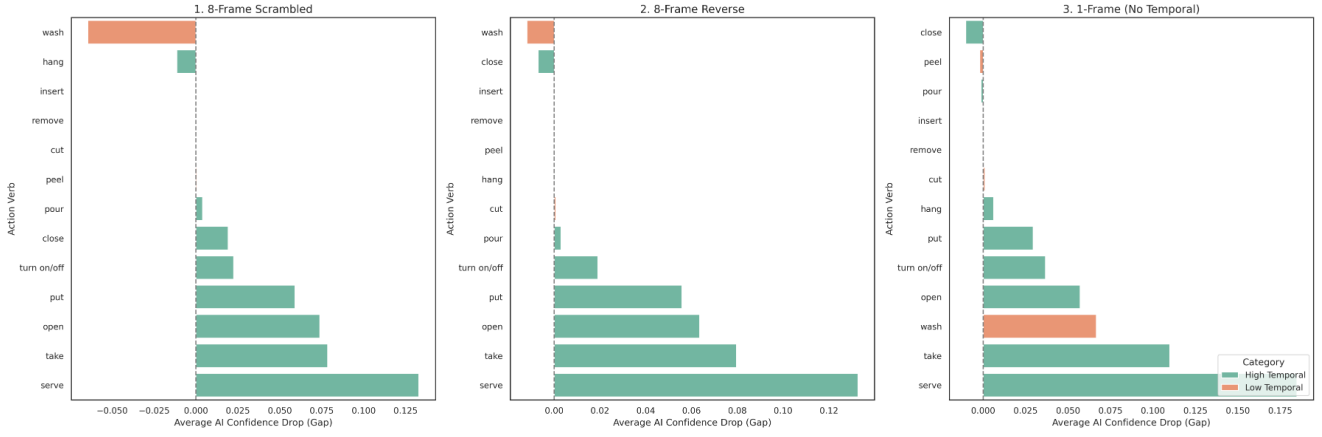


Fig. 26: Rate of gain/drop in three different settings as spatiotemporal experiment including 8-frame scrambling, reverse ordering, and 1-frame sampling.