nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01130-7

Emulating human-like adaptive vision for efficient and flexible machine visual perception

Received: 2 March 2025

Accepted: 9 September 2025

Published online: 6 November 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Human vision is highly adaptive, efficiently sampling intricate environments by sequentially fixating on task-relevant regions. In contrast, prevailing machine vision models passively process entire scenes at once, resulting in excessive resource demands scaling with spatial–temporal input resolution and model size, yielding critical limitations impeding both future advancements and real-world application. Here we introduce AdaptiveNN, a general framework aiming to enable the transition from 'passive' to 'active and adaptive' vision models. AdaptiveNN formulates visual perception as a coarse-to-fine sequential decision-making process, progressively identifying and attending to regions pertinent to the task, incrementally combining information across fixations and actively concluding observation when sufficient. We establish a theory integrating representation learning with self-rewarding reinforcement learning, enabling end-to-end training of the non-differentiable AdaptiveNN without additional supervision on fixation locations. We assess AdaptiveNN on 17 benchmarks spanning 9 tasks, including large-scale visual recognition, fine-grained discrimination, visual search, processing images from real driving and medical scenarios, language-driven embodied artificial intelligence and side-by-side comparisons with humans. AdaptiveNN achieves up to 28 times inference cost reduction without sacrificing accuracy, flexibly adapts to varying task demands and resource budgets without retraining, and provides enhanced interpretability via its fixation patterns, demonstrating a promising avenue towards efficient, flexible and interpretable computer vision. Furthermore, AdaptiveNN exhibits closely human-like perceptual behaviours in many cases, revealing its potential as a valuable tool for investigating visual cognition.

Vision is fundamental to our interpretation of the intricate physical world[1–9]. Computationally acquiring humans' visual perception capabilities is crucial for modern artificial intelligence (AI), such as multimodal large language models (MLLMs)[10–13], embodied AI agents[14–17] and medical AI[18–21]. Computer vision also carries notable implications for exploring some fundamental questions in cognitive science[22–24].

Over recent decades, machine vision models have exhibited substantial progress, approaching or surpassing expert-level performance across diverse fields, including large-scale image recognition[25–29], object detection[30], open-world visual perception[31], medical image analysis[19–21,32–34] and multimodal understanding[10–13]. However, models achieving state-of-the-art accuracy often fall short in meeting the demands of real-world applications. In real-world scenarios such as personalized AI copilots[11–13], robotic systems[14–18], wearable devices[35,36], mobile applications[37–39] and edge computing[40–42], the hardware typically faces constraints on computational capability, memory and battery capacity, yet the AI systems usually necessitate acting in real-time and performing low-latency interactions with users and

e-mail: shijis@mail.tsinghua.edu.cn; gaohuang@tsinghua.edu.cn

https://doi.org/10.1038/s42256-025-01130-7

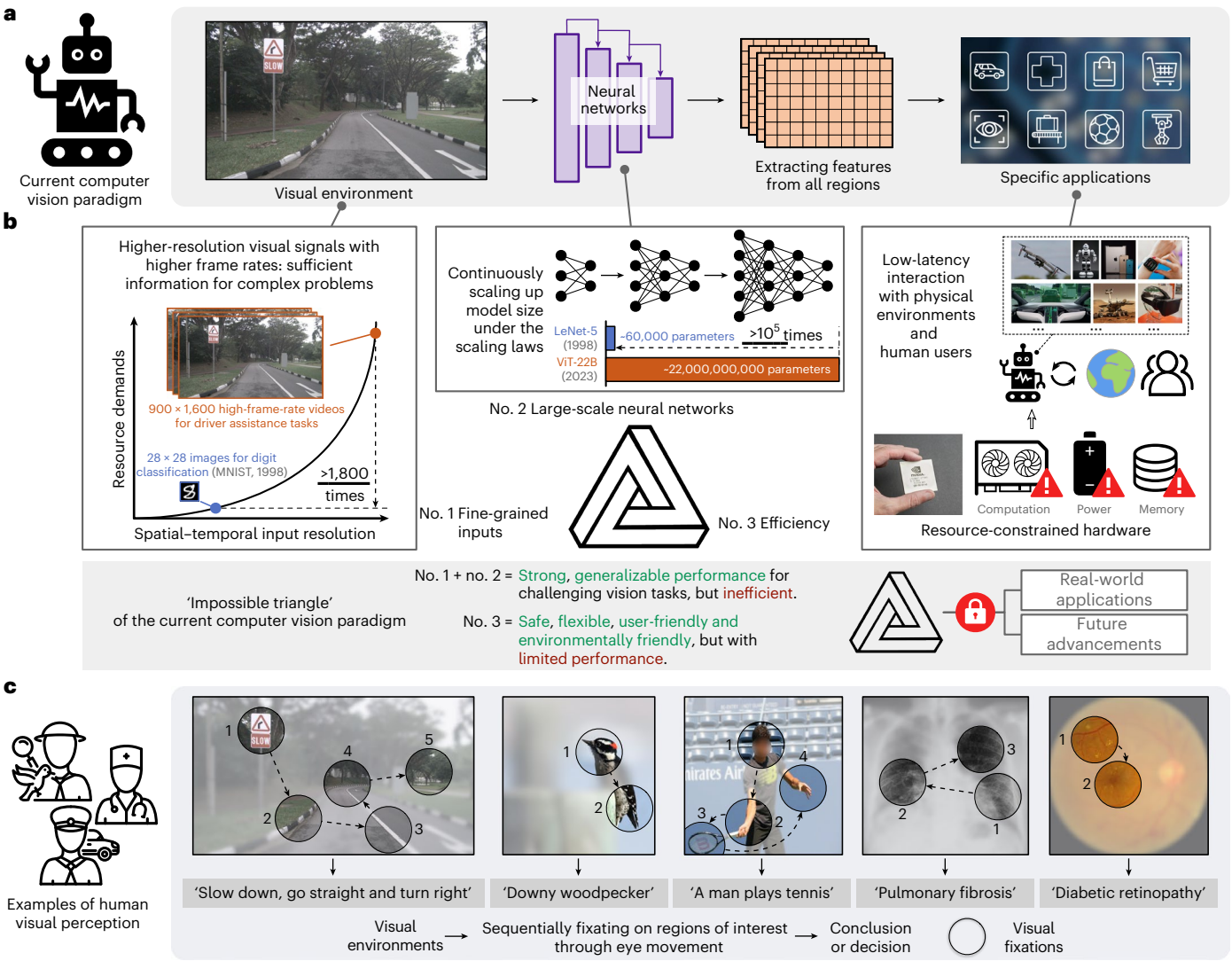


Fig. 1 | The impossible triangle faced by the current framework of machine visual perception. a, The current prevailing framework established decades ago[44–47]: a model processes the whole image in its entirety at once, with all pixels fed into neural networks simultaneously and processed parallelly, extracting features from all regions for downstream applications. All regions are equivalent in computation. **b**, However, an impossible triangle has emerged under this approach, which impedes both future advancements and adoption in diverse real-world scenarios. Specifically, continuously scaling up model size and input complexity (for example, spatial–temporal resolutions) yields superior capabilities for addressing challenging real-world vision tasks, but usually compromises efficiency, leading to dramatically growing resource demands. **c**, The human visual system circumvents this impossible triangle by using an active and adaptive perception strategy, which does not process everything everywhere all at once. Instead, human vision only acquires information when and where it is needed, which is implemented by sequentially sampling the optic array, progressively directing a high-resolution fovea towards a few regions of interest through eye movement, until the observation is sufficient. Images from: nuScenes, ref. 108 under a Creative Commons license CC BY-SA 4.0; ref. 109 under a Creative Commons license CC0 1.0; ref. 110 under a Creative Commons license CC BY-SA 4.0; ref. 111 under a Creative Commons license CC BY-SA 4.0; ref. 112 under a Creative Commons license CC0 1.0; Pixabay under a Creative Commons license CC0 1.0; Pexels under a Creative Commons license CC 1.0.

physical environments. In contrast, the inference of large-scale vision models usually involves activating millions or billions of parameters to process high-resolution images with high frame rates, leading to tremendous power consumption, considerable graphical processing unit memory requirements and non-trivial time delays. These limitations make it challenging to deploy highly capable, scaled-up models in real systems and may even pose a risk to human life by making high-latency decisions in safety-critical domains such as autonomous driving and medical robots. Besides, inference requests of computationally intensive models ultimately translate to carbon emissions, which should be minimized for environmental sustainability[43].

A foundational source of these inefficiencies is rooted in a current prevalent routine stemming from straightforwardly extending the basic representation learning framework established decades ago[44–47]: processing a whole image or video in its entirety at once, with all pixels fed into a model simultaneously and processed parallelly, equivalent in computation (Fig. 1a). As a consequence, the model's computational complexity and memory usage scale linearly with pixel count, hence quadratically with image height or width. Historically, this posed little concern decades ago, when small neural networks with thousands of parameters were used to classify tiny images such as 28 × 28 handwritten digits[45–48]. However, this has evolved into a critical limitation in modern contexts, as current models have grown five to six orders of magnitude larger, handling increasingly complex, real-world visual data. For instance, scaling from 28 × 28 images[48] to typical web images (224 × 224, ref. 25) raises computational and memory demands by 64 times, while 900 × 1,600, a relatively small size for depicting urban driving scenes, elevates resource demands by over 1,800 times. Recent

discoveries on neural network scaling laws[11,49–51] further exacerbate this challenge, indicating that continuously scaling up model size may be essential for acquiring strong, generalizable capabilities across diverse tasks. Thus, increasing demands of higher spatial–temporal resolution for inputs, the rise of larger-scale models and the necessity of efficiency in real-world applications have formed an 'impossible triangle' (Fig. 1a), which emerges as a major bottleneck faced by the current machine vision models, and its impact is expected to further markedly intensify in the future.

This article proposes drawing inspiration from the human visual system to break through the aforementioned dilemma. When interpreting the complex surrounding environments, human vision does not process everything everywhere all at once, but uses an active and selective strategy (Fig. 1b): sequentially sampling inputs by shifting a small, high-resolution fovea towards a few regions or objects of interest, and constructing a perception of the environment by combining information from different fixations over time[1,3,5,7–9]. This evolved system efficiently filters pertinent signals from extraneous information[6,52–54], markedly diminishing processing complexity[55,56]. Ultimately, regardless of the original visual environment's complexity, the resource demands of human vision depend primarily on the bandwidth and quantity of fixations: the former has been 'predefined' as a proper size, while the latter can be minimized by only acquiring information when and where it is essential for specific tasks. Thus, the human visual system incorporates tremendous numbers of neurons and demonstrates remarkable capabilities, but can efficiently handle complex real-world scenes without encountering the impossible triangle limitation (Fig. 1a) faced by modern computer vision models.

As early as 2015, ref. 57 famously argued that future computer vision systems in AI are expected to attain much progress by emulating human vision to sequentially and actively decide where to look in an intelligent, task-specific way ('The future of deep learning' in ref. 57). However, nearly a decade later, the potential of such adaptive visual systems has not yet received adequate attention. Existing studies[58–65] tend to offer limited modelling of human-like adaptiveness, typically achieving only modest efficiency gains, restricted to small datasets or specific tasks and lacking comprehensive, theoretically grounded frameworks applicable across diverse architectures and tasks. See Supplementary Section 1 for discussions on related studies. These limitations need to be urgently addressed and it is important to demonstrate how machine vision models can leverage the human-like adaptive perception approach to overcome the inherent effectiveness–efficiency trade-off dilemma.

In response to the pressing needs, we develop an AdaptiveNN framework, aiming to drive a major shift from 'passive' to 'active and adaptive' vision models. AdaptiveNN formulates visual perception as a coarse-to-fine sequential decision-making process, progressively identifying and fixating on regions pertinent to the task of interest, incrementally combining information across fixations, and actively concluding its observation when sufficient information is gathered. Hence, akin to human vision, large models can be used for superior capabilities, yet their inference remains low cost since they only process a minimally necessary subset of regions within the complex scenes. We introduce a theoretical analysis integrating representation learning with self-rewarding reinforcement learning, which enables training the non-differentiable AdaptiveNN in end-to-end without relying on specialized task structures or additional annotations beyond standard objectives. Our comprehensive experimental evaluation reveals that AdaptiveNN demonstrates markedly improved efficiency, flexibility and interpretability in diverse scenarios. For example, AdaptiveNN reduces the inference cost of well-performing models by up to 28 times without sacrificing accuracy, especially effective for processing complicated real-world scenes and for using large models. It also exhibits marked behavioural flexibility to adapt to varying task instructions and fluctuating resource availability without retraining, and achieves strong interpretability through analysing its fixation patterns. These features align with the recognized advantages of human visual systems[7,52–56,58,66]. Furthermore, the perceptual behaviours of AdaptiveNN are indistinguishable from people in many cases, uncovering its potential as a useful instrument for investigating human visual cognition.

AdaptiveNN

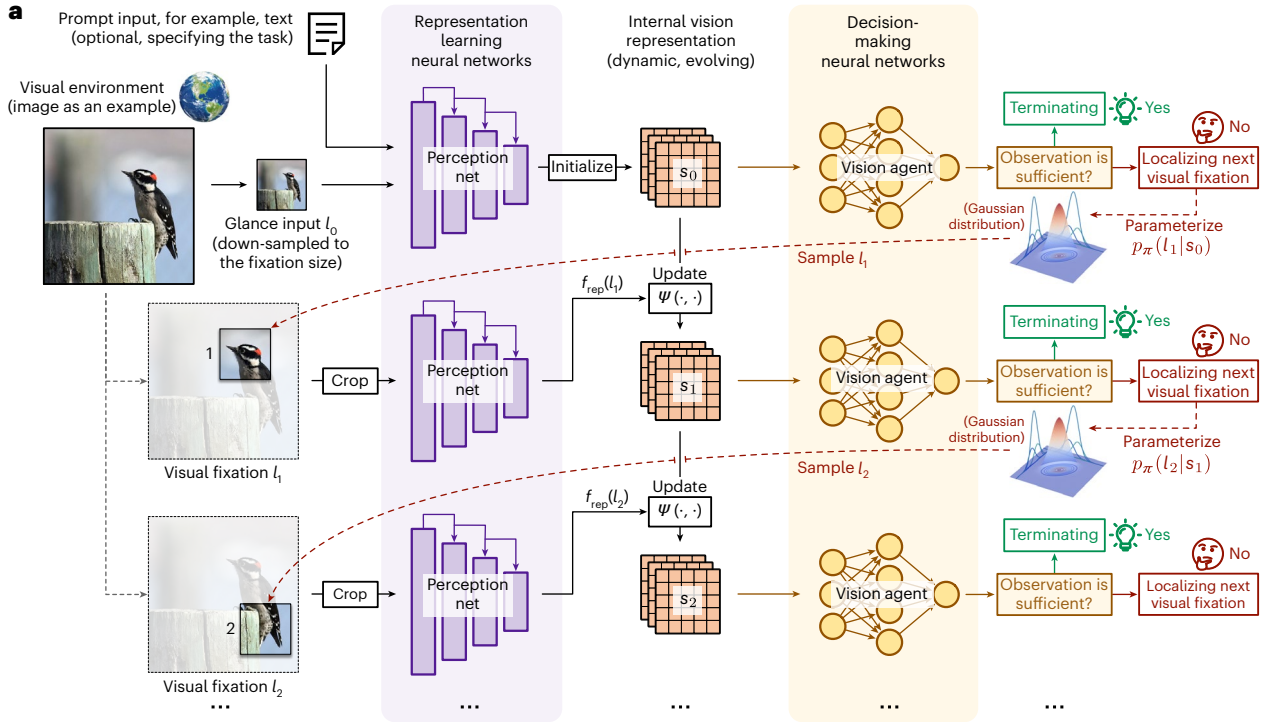
Here we briefly introduce AdaptiveNN, with more details deferred to the sections 'Architecture of AdaptiveNN' and 'Theoretical learning principles of AdaptiveNN'.

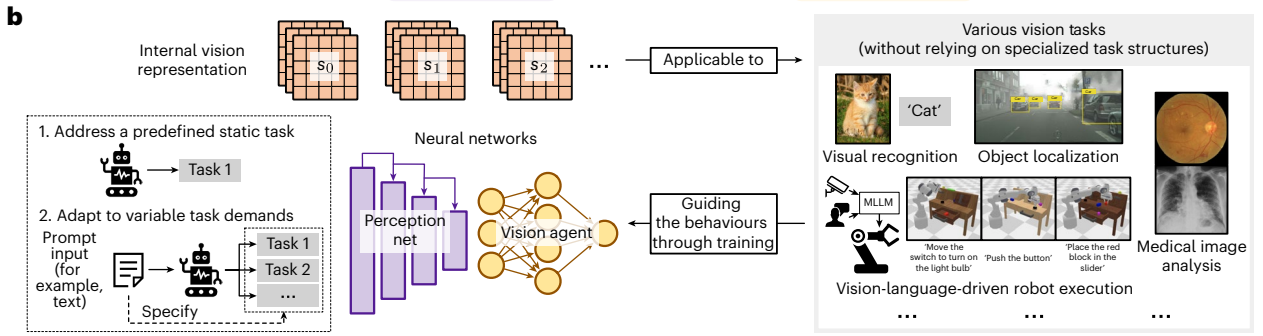
Framework

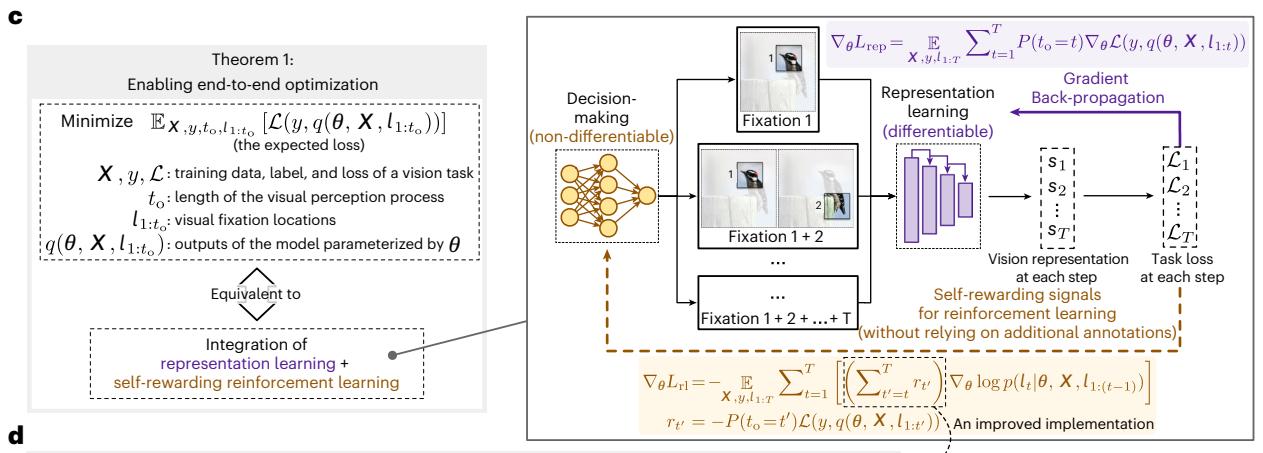
As shown in Fig. 2a, consider a generic visual environment X (for example, an image or video frame) structured as an $H \times W$ scene (where H denotes height and W width). AdaptiveNN formulates visual perception as a multi-step dynamic decision-making process, sequentially fixating on the regions of interest (denoted by $l_1, \ldots, l_t$), incrementally combining information across fixations to build up a continuously updated internal representation (denoted by $s_1, \ldots, s_t$), and actively determining when to conclude observation. At the tth step, the vision agent processes the current composite vision representation s_t, and determines whether the observation is sufficient enough to be terminated on the basis of the summarized information of previous steps and the task demands. If more information needs to be acquired from the environment, the vision agent will select the next location to fixate on through a policy network π: $l_{t+1} \sim p_\pi(l_{t+1}|s_t)$. Conversely, the perception process will not proceed, with s_t leveraged to address the task of interest. Each selected visual fixation l_t is processed by a high-capacity representation learning neural network (perception net $f_{rep}(\cdot)$) for extracting discriminative deep features to update the internal vision representation. Furthermore, the full sequential process initiates with a quick glance, where a network coarsely processes an unknown scene in a down-sampled scale to establish an initial representation. This design is inspired by the prominent theory that human vision operates in a global-to-local, coarse-to-fine manner[67–72], where humans' initial perception (vision at a glance[69]) matches a high-level, generalized, abstract scene interpretation, while later vision guides serial eye movements to attend to low-level, specific, fine receptive fields, incorporating the detailed

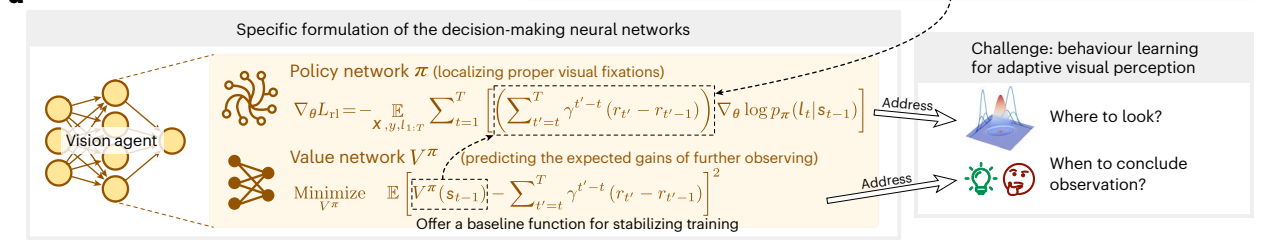
Fig. 2 | Schematic overview of AdaptiveNN. a, The architecture and inference procedure of AdaptiveNN. The model iteratively identifies new valuable regions to fixate on, and actively determines the appropriate time to conclude its observation. The information from all processed fixations is incrementally combined, forming a dynamic, evolving internal vision representation. The full sequential perception procedure initiates with a quick glance at the visual environment, emulating the coarse-to-fine strategy of the perception of human vision[67–72]. **b**, AdaptiveNN is compatible with a broad range of vision tasks, including both predefined static tasks and tasks with variable demands specified by prompt inputs (for example, text). The behaviours of AdaptiveNN are learned under task-driven supervision signals. **c**, The training of AdaptiveNN is challenging as it incorporates both continuous and discrete optimization. We address this by developing a theoretical analysis that decomposes the expected loss into an integration of representation learning and self-rewarding reinforcement learning objectives. Our method enables training AdaptiveNN in end-to-end without relying on specialized task formats or additional annotations beyond standard objectives. **d**, In implementation, we formulate the vision agent as the combination of a policy network π and a value network V^π, which addresses 'where to look' and 'when to conclude observation' simultaneously, and also facilitates a stabilized reinforcement learning process. Images adapted from ref. 78 under an MIT Licence; ref. 109 under a Creative Commons license CC0 1.0; ref. 110 under a Creative Commons license CC BY-SA 4.0; ref. 112 under a Creative Commons license CC0 1.0; Pixabay under a Creative Commons license CC0 1.0; Pexels under a Creative Commons license CC0 1.0.

https://doi.org/10.1038/s42256-025-01130-7









information available there into the perception. The behaviours of the full AdaptiveNN framework is guided through training (Fig. 2b), where it can either learn to address a predefined static task or learn to adapt to variable task demands according to a prompt input (for example, language).

In short, mimicking human vision, AdaptiveNN observes a complex visual environment through iteratively localizing and processing visual fixations, and actively deciding when its knowledge about the scene is adequate for fulfilling the given task. It focuses resources selectively on some important parts of the visual environment captured by several fixations, whose number is dynamically adjusted depending on the difficulty of accomplishing the task on top of each specific sample. The resource demands of its inference process are independent of the size, or complexity, of the visual environment to perceive. Hence, compared with the current prevailing approach that processes the full visual environment all at once, AdaptiveNN enables preserving the superior accuracy of large-scale neural networks with high-resolution inputs, but remains low-cost during inference by strategically selecting 'where to look', thus minimally suffering from the effectiveness–efficiency trade-off dilemma in previous methods.

AdaptiveNN is also appealing in allowing conveniently adjusting its average inference cost online (by varying the statistical distributions of fixation counts) without necessitating additional training, balancing efficiency–effectiveness favourably across a wide range. This enables AdaptiveNN to dynamically make full use of all available resources or obtain the required performance with minimal power consumption.

Furthermore, AdaptiveNN's formulation is general and flexible. Various off-the-shelf network architectures, such as Transformers and convolutional networks, can be readily deployed as its feature-extraction module. Moreover, the internal vision representation of AdaptiveNN does not adopt a strong assumption on its application scenarios, and may be implemented under diverse task settings, for example, using AdaptiveNN as stand-alone perceptual models or as the basis of MLLMs, being applied to static images and videos, or interacting with dynamic environments such as for robotics.

Training

Training AdaptiveNN incorporates both continuous (for example, extracting features from visual fixations) and discrete (for example, learning to select fixation positions) optimization. This cannot be solved by standard algorithms such as gradient back-propagation. To address this challenge, we present a theory enabling training AdaptiveNN in end-to-end (Fig. 2c). We prove that when considering optimizing perceptual behaviour distributions for an arbitrary vision task, an integrated formulation of representation learning and self-rewarding reinforcement learning naturally emerges as a major learning principle. The former and latter solve continuous and discrete optimization problems, respectively. Notably, this procedure does not rely on specialized task formats or additional annotations beyond the task objective itself.

Given AdaptiveNN parameterized by θ and a visual environment X, we denote the distribution of the locations of visual fixations $l_1, \ldots, l_t$ as $p(l_{1:t}|\theta, X)$. Built on this, given a vision task, the model's outputs at tth step for accomplishing the task (stemming from the internal representation s_t) are defined as $q(\theta, X, l_{1:t})$, for example, classification logits. Then, for a label y associated with X, which is defined on the task, assume that we have a performance measure (typically a loss function) $\mathcal{L}(y, q(\theta, X, l_{1:t}))$, such as the cross-entropy loss for classification and the mean squared error for regression. Consider an expected form of optimization objective:

$$\text{Minimize} \quad L(\theta) = \mathbb{E}_{X,y,t_o\sim p(t_o)} \int_{l_{1:t_o}} p(l_{1:t_o}|\theta, X)\mathcal{L}(y, q(\theta, X, l_{1:t_o})). \tag{1}$$

We have the following theorem.

Theorem 1. *The gradients of $L(\theta)$ can be decomposed into a combination of representation learning and self-rewarding reinforcement learning objectives*:

$$\nabla_\theta L(\theta) = \nabla_\theta L_{\text{rep}}(\theta) + \nabla_\theta L_{\text{rl}}(\theta), \tag{2}$$

where

$$\begin{aligned}
\nabla_\theta L_{\text{rep}} &= \underbrace{\mathbb{E}_{X,y,l_{1:T}} \sum_{t=1}^{T} P(t_o = t)\nabla_\theta \mathcal{L}(y, q(\theta, X, l_{1:t}))}_{\text{Representation learning}}, \\
\nabla_\theta L_{\text{rl}} &= \underbrace{-\mathbb{E}_{X,y,l_{1:T}} \sum_{t=1}^{T} \left[\left(\sum_{t'=t}^{T} r_{t'}\right) \nabla_\theta \log p(l_t|\theta, X, l_{1:(t-1)})\right]}_{\text{Self-rewarding reinforcement learning}}, \\
r_{t'} &= -P(t_o = t')\mathcal{L}(y, q(\theta, X, l_{1:t'})).
\end{aligned} \tag{3}$$

Proof.. Please refer to the section 'Theoretical learning principles of AdaptiveNN'.

Results

We comprehensively evaluate AdaptiveNN on 17 benchmarks organized into 9 different tasks, aiming to arrive at a complete picture of the characteristics, efficacy and potential values of AdaptiveNN. The section 'Evaluation tasks for AdaptiveNN' details these tasks.

Large-scale real-world visual understanding

We first consider the visual recognition task on ImageNet[25], which comprises >1.28 million images in 1,000 categories (including various objects, buildings, humans, animals, scenes and so on), and is widely acknowledged for its critical role in assessing machine learning methods[26–28]. We deploy two representative feature-extraction backbones within AdaptiveNN to demonstrate its generalizability, namely ResNet (convolutional network)[26] and DeiT (vision transformer)[73], each representing a wide range of popular architectures.

Figure 3a illustrates the visual perception behaviours learned by AdaptiveNN. AdaptiveNN consistently fixates on the class-discriminative regions, such as animals' heads, musical instruments' principal structures and coffee machines' functional parts such as knobs and nozzles. Moreover, for complex and atypical visual inputs, AdaptiveNN adjusts by extending its observation duration to enhance prediction accuracy. This adaptive behaviour is particularly evident when the objects of interest are small, located distant from the camera or depicted from uncommon perspectives, showcasing only parts of their entirety.

Quantitatively, emulating human-like adaptiveness in vision models substantially enhances computational efficiency and adaptability (Fig. 3b,c and Supplementary Tables 6 and 9). On top of the same backbones, AdaptiveNN-DeiT-S and AdaptiveNN-ResNet-50 achieve accuracies on par with their traditional, non-adaptive counterparts (81.6% and 79.1%) at the computational costs of 2.86 and 3.37 GFLOPs per image, reflecting efficiency gains of 5.4× and 3.6×, respectively. In addition, AdaptiveNN's computational cost can be adjusted online flexibly, yielding a favourable efficiency–effectiveness balance across broad ranges, while non-adaptive models typically require retraining to achieve similar adjustments. Figure 3d and Supplementary Tables 10 and 11 show that on average, leveraging progressively more fixations improves accuracy significantly (all $P < 0.005$). Figure 3e and Supplementary Tables 12 and 13 illustrate that AdaptiveNN adapts to dynamic resource constraints by allocating more fixations to relatively difficult examples and fewer to simpler ones, optimizing overall resource distributions to maximize performance efficiency under variable computational budgets.

Fine-grained visual recognition

We further evaluate AdaptiveNN on six fine-grained recognition tasks, where minor inter-class differences and high intra-class variability necessitate precise localization of subtle, task-specific signals amid

https://doi.org/10.1038/s42256-025-01130-7

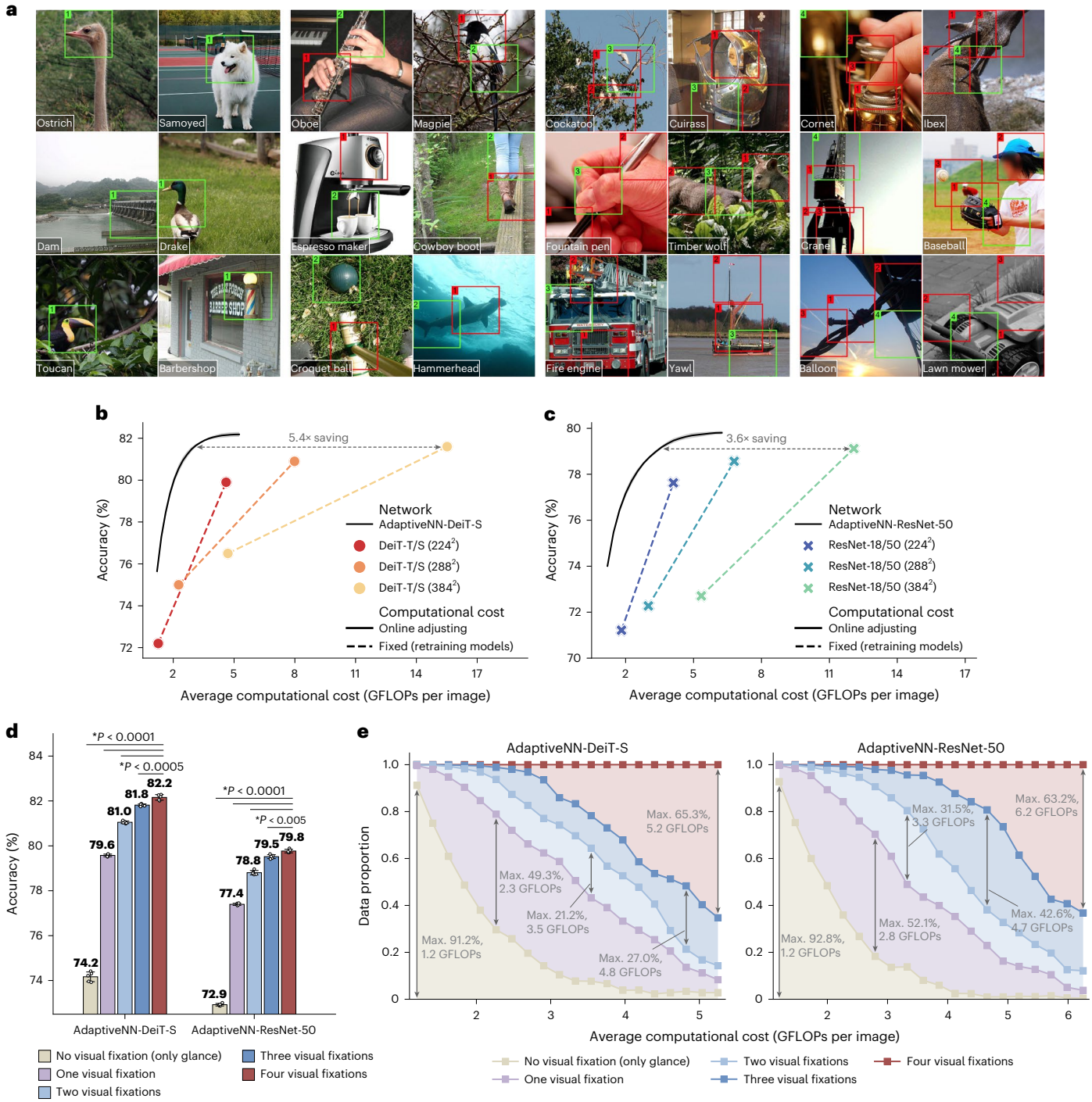


Fig. 3 | Results of ImageNet large-scale real-world visual understanding. **a**, Qualitative assessment showcasing the visual fixations localized by AdaptiveNN(-DeiT-S), with boxes marking the locations of fixations and colours indicating the model's decision to conclude (green) or continue (red) observation at each step. Step indices are presented at the top left of the boxes. Ground-truth labels are displayed at the bottom left of the images. **b**,**c**, Quantitative comparisons of AdaptiveNN(-DeiT-S) (**b**) and AdaptiveNN(-ResNet-50) (**c**) and traditional non-adaptive models on top of identical backbones: ImageNet-1K top-1 validation accuracy versus average computational cost for inferring the model. To obtain non-adaptive models with varying costs, we consider two common approaches: adjusting model sizes and input resolutions. **d**, Relationship between validation accuracy and the number of visual fixations, assuming that all samples use the same number of visual fixations. Exact P (>0.0001) values: 0.00045, 0.0014. **e**, Proportions of data that use different numbers of visual fixations, set against different budget constraints for computational costs. In **b**–**d**, the results show means ± standard deviations from five independent trials with different random seeds. *One-way analysis of variance with Tukey's honestly significant difference test. Max., maximum. Images adapted from ref. 113 under a Creative Commons license CC0 1.0.

overwhelming irrelevant information. This probes into AdaptiveNN's potential in mirroring human vision's key strength in nuanced perceptual capabilities[7,53–56,58,66]. Extended Data Fig. 1a and Supplementary Tables 14 and 19 present quantitative results. AdaptiveNN achieves substantial computational savings (6.2×, 6.1×, 7.6×, 8.2×, 5.8×, 6.3×) without sacrificing accuracy, underscoring our model's human-like proficiency in fixating on and leveraging fine-grained discriminative features. Similar to ImageNet, AdaptiveNN demonstrates good interpretability. As

https://doi.org/10.1038/s42256-025-01130-7

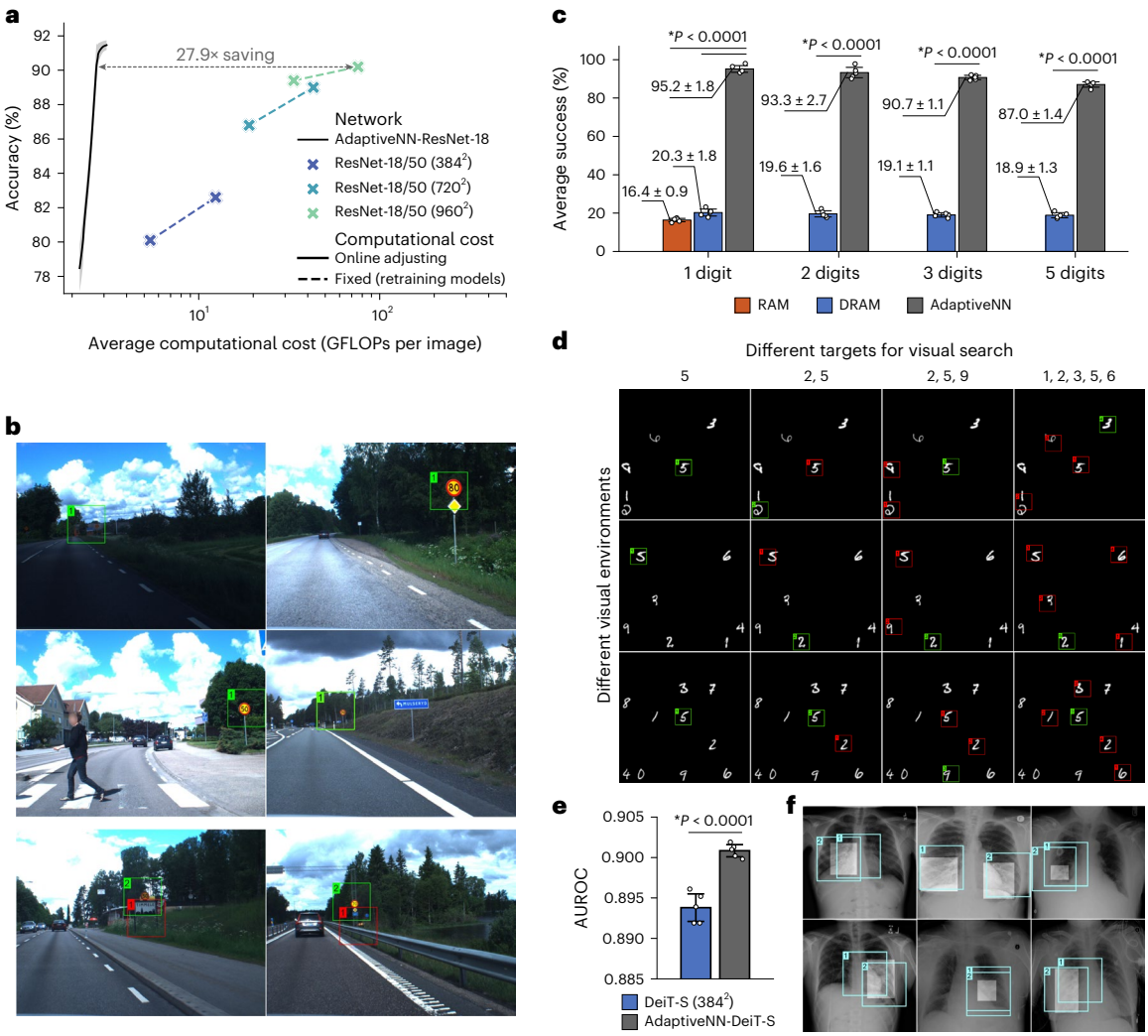


Fig. 4 | Assessment of AdaptiveNN in more general visual perception scenarios: processing images from real driving and medical scenarios, and visual search tasks with variable demands. a, Comparisons of AdaptiveNN and conventional non-adaptive models in processing complicated, non-object-centric real-world scenes: top-1 validation accuracy versus average computational cost for inferring the model (log-scale). We consider the traffic sign recognition task on the STSD[74], composed of 960 × 1,280 road-scene images collected on real moving vehicles. ResNets are deployed as backbones since convolutional networks tend to be more efficient for processing high-resolution inputs. **c**, Average success rates of visual search tasks. Here '*n* digits' indicates the number of target digits, while the bars indicate the means ± standard deviations of five randomly generated visual search tasks with various target categories (yet maintaining a constant number of targets). Success is defined as accurately retrieving exactly all the digits specified by a given task. **e**, AUROC of the RSNA pneumonia detection task[76]. All models are trained to predict the presence or absence of pneumonia on the basis of image-level labels. Here we do not perform adaptive termination in AdaptiveNN and mainly focus on the AUROC after processing all fixations, since efficiency may not be a major focus of medical diagnosis tasks. **b**,**d**,**f**, Qualitative evaluation results corresponding to **a**, **c** and **e**, respectively. Boxes represent visual-fixation locations, with colours indicating the model's decision to either continue (red) or terminate (green) observation at that step. Step indices are annotated at the upper left corner of each box. **f**, Lighter boxes show the pneumonia regions annotated by human clinicians (this localization information is not used for training). Except for **c**, the results show means ± standard deviations from five independent trials with different random seeds. *Two-sided independent samples *t*-test. Data from ref. 76. Images adapted from ref. 74, Springer Nature Limited.

shown in Extended Data Fig. 1b–e, it autonomously learns to focus on critical clues (for example, bird beaks, car lights, airplane engines and propellers) without explicit localization supervision.

Efficient processing of visual data from real driving scenarios

ImageNet and fine-grained recognition datasets often feature object-centred images curated by humans. Nevertheless, AdaptiveNN, like human vision, is applicable to more general and complex scenarios, for example, efficiently processing non-object-centric images collected in the wild without specified preprocessing. To demonstrate this, we evaluate AdaptiveNN on the Swedish traffic signs dataset (STSD)[74], comprising high-resolution road-scene images collected on real moving vehicles. The objects of interest are very small, distributed diversely, and not clear in many cases, presenting realistic challenges.

Illustrated in Fig. 4a and Supplementary Tables 20 and 21, AdaptiveNN matches non-adaptive ResNet-50's 90.2% accuracy while reducing computational cost from ~76 GFLOPs per image to ~2.7 GFLOPs per image (27.9× reduction). This substantial enhancement can be elucidated through qualitative analysis (Fig. 4b). AdaptiveNN's fixations adaptively focus on the small, task-relevant regions within the expansive, intricate and cluttered visual scenes, mirroring the efficiency characteristic of human vision. Moreover, AdaptiveNN can recognize and correct initial misidentifications by refining its fixations on the basis of contextual cues.

Addressing vision tasks with flexible requirements

Humans can flexibly adjust perception behaviours, such as fixation locations and counts, conditioned on diversified, task-specific

demands[7,9,53]. To examine whether AdaptiveNN exhibits similar adaptability, we designed a visual search scenario with 224^2 images containing 6–10 non-repeating digits on a black background. A model is trained to identify the locations of specified digits, with the setups of target categories and numbers flexibly varied, each constituting an individual visual search task.

Figure 4c and Supplementary Table 22 summarize the expected success rate of correctly retrieving all specified targets across diverse tasks and visual environments. AdaptiveNN consistently achieves ~90% success rates, regardless of target counts. By contrast, existing popular human-like sequential perception models (RAM[58], DRAM[59]) remain below ~20%, underperforming AdaptiveNN by ~4.5×. Figure 4d illustrates AdaptiveNN's capability to adaptively modulate its fixation selection and observation termination strategies conditioned on each specific search task and input.

Interpretability-critical tasks: image processing in medical scenarios

Similar to understanding human vision[52,66,75], AdaptiveNN's fixation patterns offer a critical window into interpreting its decision-making processes (Figs. 3a and 4b–d and Extended Data Fig. 1b–e). Leveraging this insight, we evaluate its performance in interpretability-critical tasks, exemplified by pneumonia detection from chest X-rays[76]. Trained only on image-level pneumonia labels, AdaptiveNN achieves significantly superior area under the receiver operating characteristic curve (AUROC) on validation data than non-adaptive models ($P < 0.0001$; Fig. 4e). Notably, despite no explicit localization supervision, AdaptiveNN's fixations (Fig. 4f) closely match the pulmonary opacity regions annotated by 18 board-certified radiologists from 16 institutions. This concordance uncovers AdaptiveNN's potential in AI applications demanding not only precision but also interpretability, such as healthcare applications.

Embodied MLLMs based on AdaptiveNN

The formulation of AdaptiveNN is sufficiently general to be deployed as the perceptual module of an embodied agent interacting with dynamic physical environments. We integrate AdaptiveNN into a MLLM based on RoboFlamingo[77] (Extended Data Fig. 2a). The MLLM receives language prompts, observes the environment (with or without AdaptiveNN) and updates a recurrent policy network to execute actions, iterating through observation–action cycles (Fig. 5a). We use the CALVIN LH-MTLC benchmark[78], where agents need to complete five-subtask sequences described in natural language and the average successful length (0–5) across 1,000 sequences is evaluated (Extended Data Fig. 2b). We consider two settings using identical validation tasks and different training data scales (that is, D → D, ABCD → D). Figure 5b and Supplementary Tables 23 and 24 demonstrate that AdaptiveNN reduces computational cost by 4.4–5.9× without sacrificing effectiveness, and is notably more flexible in adjusting its resource demands online without retraining. Success rates across task types are detailed in Supplementary Tables 25 and 26. Figure 5c and Supplementary Tables 27 and 28 report the average performance corresponding to each fixed count of visual fixations, depicting a progressively increasing trend, more pronounced for large-scale, diverse training data such as ABCD → D. Qualitative results (Fig. 5d) show AdaptiveNN dynamically fixating on task-relevant objects and their interactions with robotic operational components, guided by visual input and language prompts. For challenging tasks such as fine-grained control, AdaptiveNN allocates more fixations for precision; otherwise, it minimizes fixations to conserve resources.

Comparisons between AdaptiveNN and human visual perception

AdaptiveNN also emerges as a potent computational tool for probing human visual cognition under controlled conditions. This can be uncovered with the marked consistency between humans and AdaptiveNN in side-by-side evaluations on the same tests of visual perception behaviours. The section 'Comparisons with human visual perception behaviours' details our experimental protocols.

First, we assess the spatial-wise consistency between human and AdaptiveNN fixations using the saliency in context (SALICON) dataset[79], where ~60 human participants freely viewed each image for 5 seconds, and their aggregated gaze density maps serve as ground truth. Figure 6b offers qualitative comparisons between the fixation regions selected by AdaptiveNN (boxes) and human gazing locations (heat maps). Our model produces human-like patterns in many cases, frequently being attracted by faces, hands, human bodies, human actions or objects intimately associated with human activity, such as food, computers, skateboards, tennis rackets and buses. Figure 6a and Supplementary Tables 29 and 30 present quantitative results. In terms of the alignment with the ground-truth spatial-adaptive human visual perception behaviours, AdaptiveNN matches or surpasses the average performance of an arbitrary individual human observer.

Second, we examine whether AdaptiveNN aligns with human judgements in assessing which visual environments are more challenging for a given task and necessitate more thorough scrutiny. Human participants ($n = 10$) rated images from six representative ImageNet categories by classification difficulty. In Fig. 6c and Supplementary Table 31, the averaged individual-normalized human-assessed scores are compared against AdaptiveNN's normalized state values, which reflect our model's assessments of each image's difficulty level. The model's estimates strongly correlate with human judgements (all $P < 0.0001$; Pearson correlation coefficient $\rho \in [0.54, 0.80]$). Figure 6d illustrates representative 'easy' and 'difficult' data identified by AdaptiveNN, where typical, clear-content images tend to be deemed 'easy'.

Finally, we establish several visual Turing tests[22] (Extended Data Fig. 3a). Human judges ($n = 39$) are given paired examples of visual perception behaviours from humans and AdaptiveNN, and instructed to identify which come from the machine. Each participant has completed 216 trials to investigate both spatial-wise visual-fixation and sample-wise difficulty-assessment behaviours. Human judgements' accuracy are evaluated: 50% indicates perfectly indistinguishable behaviours from humans, while 100% represents the worst case. We randomly replace the 'machine' behaviours of some trials with 'human' or 'random' behaviours without letting participants know, and separately evaluate the accuracy of these trials, establishing two randomized control baselines for comparison.

Results are summarized in Fig. 6e,f, Extended Data Fig. 3b,c, Supplementary Tables 32 and 33 and Supplementary Figs. 14 and 15. In all scenarios, human judges achieve only 50–51% accuracies in correctly identifying 'AdaptiveNN versus human', which do not acceptably outperform random guessing in statistics ($t(38) = 0.90, -0.09$, $P = 0.37, 0.93$). In addition, these 'machine versus human' judgements do not exhibit a significant difference from the 49–50% accuracies of the 'human versus human' baselines ($t(38) = 0.97, 0.40$, $P = 0.33, 0.69$). In contrast, 'random versus human' results in considerably easier Turing test tasks (accuracies ≥80%). These observations demonstrate that, in general, AdaptiveNN approaches an indistinguishable level from the adaptive perceptual behaviours of human vision.

Discussion

Human vision is distinguished by its remarkable flexibility to adapt to spatial regions with different content, varying complexities of visual environments, diverse task demands and fluctuating resource availability for perception. In contrast, current machine vision models mainly adopt 'passive' approaches, which usually perceive everything everywhere in parallel with an identical computational graph, regardless of the specific characteristics of variable visual environments, tasks and resources. As a consequence, high-dimensional visual inputs, large-scale neural networks and efficiency converge to an 'impossible

https://doi.org/10.1038/s42256-025-01130-7

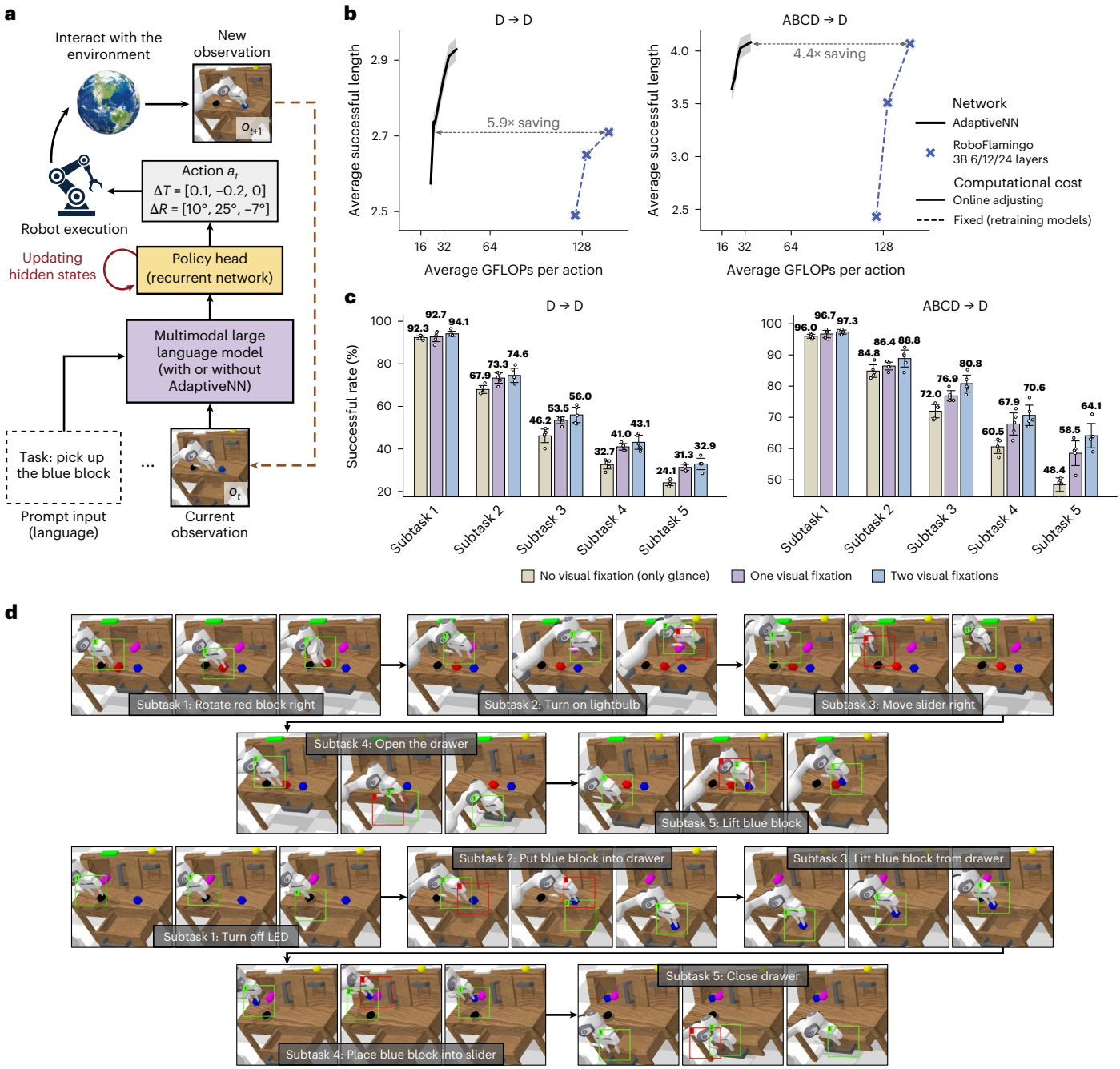


Fig. 5 | Performance of the embodied MLLMs based on AdaptiveNN. **a**, A schematic overview of the embodied MLLM. The prompt input specifies the task, and the MLLM iteratively perceives the environment to execute appropriate robotic actions, that is, six-degree-of-freedom transformation vectors in 3D space. The next observation reflects the outcome of the preceding actions. A recurrent policy head integrates information from all previous observations. **b**, Comparisons of AdaptiveNN-based MLLM and non-adaptive MLLM using identical backbones on CALVIN: average successful length (of 1,000 5-task sequences) versus average computational cost for inferring the model. For the non-adaptive models, computational costs are modulated by adjusting model sizes. D → D and ABCD → D indicate different scales of training data. **c**, Relationship between average successful rates of each subtask within task sequences and the number of visual fixations, assuming that all samples use the same number of visual fixations. **d**, Qualitative assessment of two representative five-task sequences. Boxes mark the fixation locations, and colours indicate the model's decision to conclude (green) or continue (red) observation at each step. Step indices are presented at the top left of the boxes. The prompt inputs for specifying tasks are displayed within the black boxes. In **b** and **c**, the results show mean ± s.d. from five independent trials with different random seeds. Images are constructed based on code from ref. 78.

triangle'; under the scaling laws, the first two tend to be essential for solving complex real-world vision-based problems, but they substantially compromise efficiency. This inherent limitation impedes both future advancements and application in diverse real-world scenarios. This article proposes AdaptiveNN, aiming to address this issue by enabling neural networks to emulate the adaptive behaviours of human vision, thus driving a fundamental shift in approach from 'passive' to 'active and adaptive' vision models.

Under this goal, AdaptiveNN is carefully designed to be general, compatible with various network architectures and tasks. Extensive evaluations uncover that AdaptiveNN reduces the computational cost of well-performing vision models by up to 28× without sacrificing

accuracy. Moreover, AdaptiveNN exhibits human-like flexibility in adjusting its resource demands online without necessitating additional training, as well as in customizing its perceptual strategies conditioned on variable task demands through modifying training objectives or introducing language prompts as inputs. Besides, AdaptiveNN is distinctive in its enhanced interpretability through analysing its fixation patterns, in a manner akin to understanding human visual systems[52,75,80,81]. We believe these superiorities demonstrate a practical avenue towards the next generation of energy-efficient, flexible and interpretable machine visual perception frameworks.

AdaptiveNN may also offer valuable insights into equipping computer vision models with adaptive sequential 'reasoning'-like perception capabilities using reinforcement learning, analogous to the approach employed in DeepSeek-R1 (ref. 82). We demonstrate how to model visual perception tasks as sequential decision procedures, and reveal why and how such models should be trained using reinforcement learning. The resulting models can adaptively use a larger number of strategically selected visual fixations to handle more challenging vision tasks.

In addition, AdaptiveNN potentially emerges as a useful computational instrument for advancing the understanding of human behavioural and learning processes. For example, AdaptiveNN, learned solely on large-scale, object-centric visual recognition tasks, exhibits indistinguishable behaviours from human vision in many cases, in terms of either the 'eye movement' patterns in novel scene observation or assessing the 'difficulty levels' of various visual environments. Hence, highly human-like behaviours of actively observing objects and scenes may be learnable through routine vision tasks such as recognition, without the guidance of strong innate inductive biases (for example, biases concerning objects, agents, space and biological motion[83–90]). In this sense, we also[23,24,91,92] propose to leverage advanced AI methods such as deep networks and reinforcement learning to explore fundamental cognitive science questions under controlled experimental conditions. The potential of this interdisciplinary exploration is further discussed in Supplementary Section 2.

Methods

Architecture of AdaptiveNN

Here we describe the major components of AdaptiveNN (Fig. 2a) in detail. More implementation details on the architectures of AdaptiveNN can be found in Supplementary Sections 4.2 and 4.3.

Visual fixations. $l_1, \ldots, l_t$ (at the first, …, tth steps). AdaptiveNN never senses the visual environment in its entirety. In contrast, it extracts information from a sequence of smaller, bandwidth-limited inputs corresponding to certain local regions of the environment, named visual fixations, denoted by $l_1, \ldots, l_t$. AdaptiveNN actively determines the locations of $l_1, \ldots, l_t$ step by step, under the goal of maximizing their contributions to the task of interest, until sufficient information has been gathered. The small bandwidth of visual fixations ensures that the resource demands of AdaptiveNN can be controlled independently of the size or complexity of the original visual environments, and will not grow dramatically with higher spatial–temporal input resolution. As a consequence, visual perception can be efficient even when using large-scale neural networks to perceive intricate real-world scenes with high frame rates. Furthermore, since the fixations are strategically localized to focus on the important visual content and new fixations will be continuously introduced until the observation is sufficient, the model performance can be maximally preserved. In some scenarios, the performance may even be improved by eliminating task-irrelevant information interference. Without loss of generality, we define a visual fixation as a $P \times P$ patch ($P < H, W$) to be compatible with most modern deep learning scenarios[26–28,57]. Although we consider the most general form of square patches as fixations to ensure the generality of our framework, more advanced fixation formats may be adopted for optimization towards specific models of tasks (for example, multi-scale mixed visual fixations).

Perception net. f_{rep} is a representation learning backbone network that converts raw pixelated image inputs into deep representations with semantic meanings. As mentioned before, high-capacity, large-scale models can be used as f_{rep}, to obtain strong visual processing capabilities. Since f_{rep} only needs to process the bandwidth-limited visual fixation, its inference still enjoys superior efficiency.

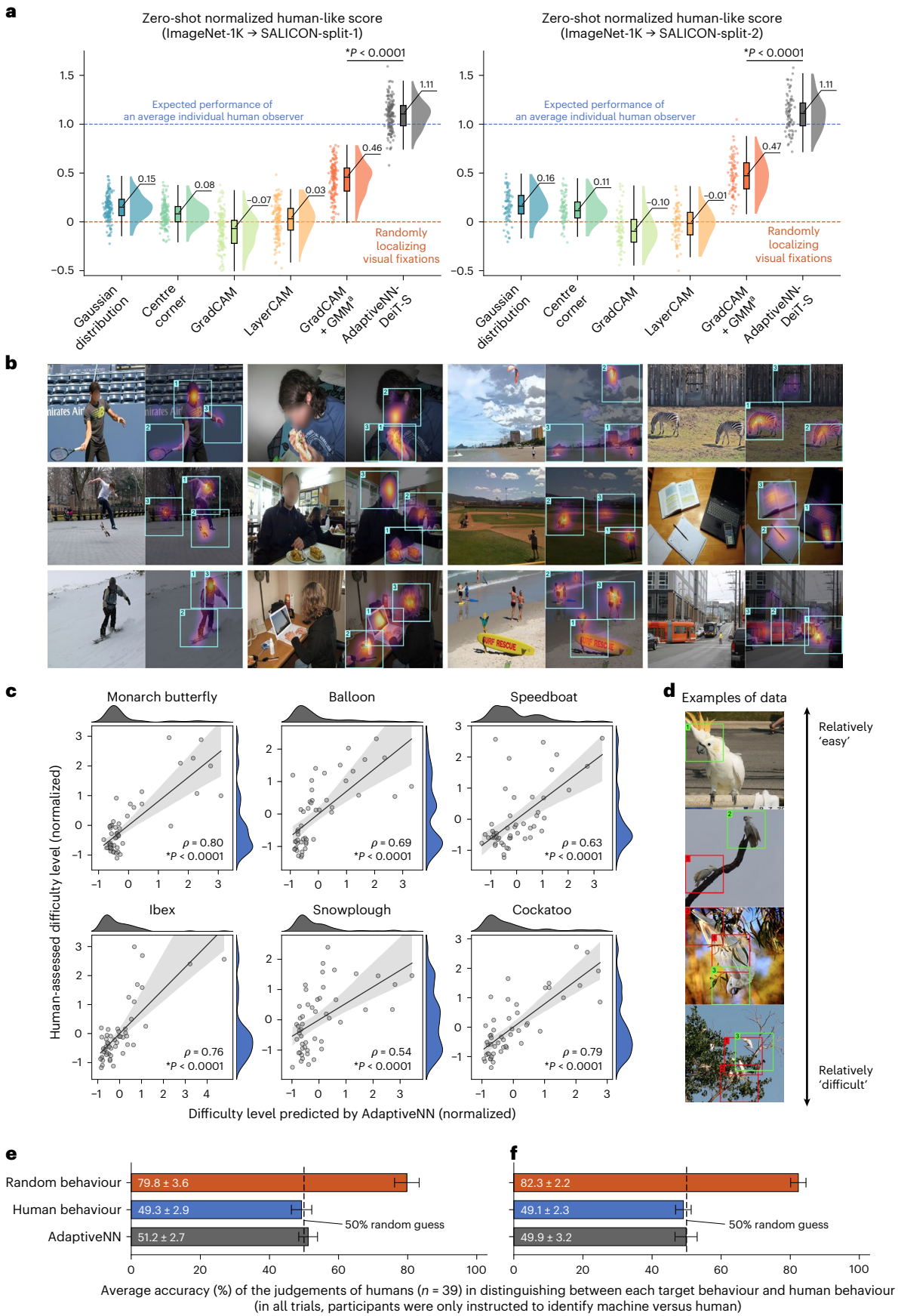
Internal vision representation. $s_1, \ldots, s_t$ is maintained during the whole visual perception process, and dynamically updated using the features extracted from each visual fixation by f_{rep}, namely

$$s_t = \Psi(s_{t-1}, f_{\text{rep}}(l_t)), \tag{4}$$

where $\Psi(\cdot, \cdot)$ denotes the updating operator (see Supplementary Section 4.2 for its implementation details). The internal representation s_t summarizes the information from the history of all past observations, encoding the model's current knowledge of the environment. It serves two critical purposes. First, as shown in Fig. 2b, s_t is the output of the AdaptiveNN framework, and the information within it will be used to fulfil the given vision task (feeding s_t into a task-specific head, detailed

Fig. 6 | Behavioural comparisons between AdaptiveNN and human vision. **a**, Normalized human-like scores, which quantify the probability that the ground-truth gazing centres of human vision (whose distribution is estimated by averaging across ~60 observers' visual perception behaviours) fall into the visual-fixation regions localized by AdaptiveNN or comparative strategies. The raw results are normalized with respect to the expected performance of selecting fixation regions with the gazing locations of an individual human observer (1.0 on the y axis) and uniformly at random (0.0 on the y axis). Each point represents the result over a mini-batch of data (64 randomly sampled examples), while boxplots depict the distribution of results (n = 157/79 batches for SALICON-split-1/2), showing median (centre line), 25–75% percentiles (box), minima and maxima (whiskers). The evaluation is based on the SALICON dataset[79]. Our model is trained on ImageNet, having never seen the data in SALICON. This 'zero-shot' evaluation setting evaluates directly transferring AdaptiveNN's perceptual behaviours to novel, complex environments, with a fixed number of fixations, mirroring the collection procedure of human gazing centres. Baselines for comparison incorporate selecting fixation regions using (1) predefined rules, (2) class activation maps (CAM) and (3) CAM augmented with a Gaussian mixture model (GMM). See Supplementary Section 3 for the details of these baselines. *Two-sided independent samples t-test. **b**, Qualitative comparisons between the ground-truth density maps of human gazing centres (heat maps) and AdaptiveNN fixation regions (boxes). Boxes indicate fixation locations, with step indices annotated at the upper left corner of each box. **c**, Correlation of human-assessed difficulty scores (averaged across n = 10 participants) and difficulty levels (state values) evaluated by the vision agent of AdaptiveNN. Without loss of generality, the state values are taken from the first step of sequential perception processes. Results are based on six representative categories of data in the ImageNet validation set. ρ, Pearson correlation coefficients. *Two-sided correlation t-test. Error bars show the 95% confidence interval. **d**, Visualization examples of the 'easy' and 'difficult' data identified by AdaptiveNN. **e**,**f**, Results of visual Turing tests: visual fixation behaviours (**e**) and difficulty-assessment behaviours (**f**). Human judges (n = 39) are randomly given paired examples of visual perception behaviours from 'humans' and 'one within {AdaptiveNN, humans, random behaviours}'. They are instructed to identify the machine (even in cases when the pairs of 'random versus human' or 'human versus human' are given, which serve as control groups for comparison). Bars show the mean accuracy across human judges and the corresponding 95% confidence interval. Ideal performance is 50%, where the machine is indistinguishable from human behaviours in these binary choice tasks. Data points and their distributions are given in Extended Data Fig. 3b,c. [a]The GMM introduces additional computation. Images adapted from ref. 111 under a Creative Commons license CC BY-SA 4.0; ref. 113 under a Creative Commons license CC0 1.0.

https://doi.org/10.1038/s42256-025-01130-7



in Supplementary Section 4.2). Second, s_t provides necessary information for decision-making in the sequential adaptive visual perception process: that is, deciding whether to conclude observation now and where to look next. Both of these abilities are acquired through being trained to accomplish the vision task of interest (Fig. 2b).

Vision agent. The vision agent is a decision-making neural network that receives the internal vision representation $s_1, \ldots, s_t$ as inputs. At each step of the sequential perception process, it makes two decisions: assessing whether to terminate the ongoing observation and, if necessary, determining the subsequent visual-fixation location. To achieve both of them simultaneously, we formulate the vision agent as the combination of a policy network π and a value network V^π (Fig. 2d). This formulation is naturally derived from the theoretical learning principles of AdaptiveNN, which will be discussed in the section 'Theoretical learning principles of AdaptiveNN' coupled with the training algorithm of π and V^π. Here we first introduce the inference process of π and V^π. At the tth step of inference, the outputs of π parameterize a distribution from which we can sample the location of l_{t+1}, namely

$$l_{t+1} \sim p_\pi(l_{t+1}|s_t). \tag{5}$$

Paired with π, the value network V^π uses s_t to predict the expected gains of performing further observation on top of s_t (that is, further updating s_t) using π, yielding a state value $V^\pi(s_t)$. We compare $V^\pi(s_t)$ with a threshold η_t. If $V^\pi(s_t) \le \eta_t$, we are indicated that further observing is not valuable enough, and the sequential perception process will be concluded. Otherwise, $V^\pi(s_t) > \eta_t$ reveals that more fixations may yield notable improvements, and thus the new fixation l_{t+1} will be processed, evoking the $(t+1)$th step. The value of η_t is solved on the validation data, and can be adjusted online to vary the average resource demands of AdaptiveNN without additional training (Supplementary Section 4.1). Notably, the outputs of π and V^π consider both the current specific situations as the observation on each particular visual environment progresses, as well as the demands of the given vision task. The former has been encoded into s_t, while the latter is attained through the training process (section 'Theoretical learning principles of AdaptiveNN'), where π and V^π can either learn to address a predefined static task, or learn to adapt to variable task demands on top of a prompt input (for example, text), as depicted in Fig. 2b. Moreover, it is noteworthy that $V^\pi(s_t)$ reflects the model's subjective assessments, namely whether the perception process of AdaptiveNN itself is worth proceeding, while η_t determining whether $V^\pi(s_t)$ is sufficiently small represents the objective constraints imposed by the external environment, for example, the extent to which the overall available resources for visual perception are adequate in the current circumstance. This decoupled modelling of subjective and objective factors enables more flexible usage of our framework.

Theoretical learning principles of AdaptiveNN

During training, AdaptiveNN focuses on learning a model capable of sequentially attending to proper visual fixations within a complex visual environment, and extracting information from these fixations to accomplish the vision task of interest. Its optimization objective is defined as minimizing the expected performance measure of the task, namely, equation (6) (restating equation (1) for convenience).

$$\text{Minimize} \quad L(\theta) = \mathbb{E}_{X,y,t_o\sim p(t_o)} \int_{l_{1:t_o}} p(l_{1:t_o}|\theta, X)\mathcal{L}(y, q(\theta, X, l_{1:t_o})). \tag{6}$$

In equation (6), $t_o \sim p(t_o)$, $t_o \in \{1, \ldots, T\}$ indicates that during training, the total length t_o of the sequential perception process is sampled from a fixed prior distribution $p(t_o)$, which reflects the training process's statistical-level preference on the perception procedure's length. This consideration is introduced to add analytical flexibility to our model. Besides, note that we do not explicitly formulate the actions of actively concluding observation in equation (6). Conversely, we will demonstrate that the ability to evaluate when the observation is sufficient can be conveniently acquired on top of the model learned by minimizing equation (6).

Theorem 2. *(Restating Theorem 1 for convenience) The gradients of $L(\theta)$ can be decomposed into a combination of representation learning and self-rewarding reinforcement learning objectives*:

$$\nabla_\theta L(\theta) = \nabla_\theta L_{\text{rep}}(\theta) + \nabla_\theta L_{\text{rl}}(\theta), \tag{7}$$

where

$$\begin{aligned}
\nabla_\theta L_{\text{rep}} &= \underbrace{\mathbb{E}_{X,y,l_{1:T}} \textstyle\sum_{t=1}^{T} P(t_o = t)\nabla_\theta \mathcal{L}(y, q(\theta, X, l_{1:t}))}_{\text{Representation learning}} \\
\nabla_\theta L_{\text{rl}} &= \underbrace{-\mathbb{E}_{X,y,l_{1:T}} \textstyle\sum_{t=1}^{T} \left[\left(\sum_{t'=t}^{T} r_{t'}\right) \nabla_\theta \log p(l_t|\theta, X, l_{1:(t-1)})\right]}_{\text{Self-rewarding reinforcement learning}}, \\
r_{t'} &= -P(t_o = t')\mathcal{L}(y, q(\theta, X, l_{1:t'})).
\end{aligned} \tag{8}$$

Proof.. Taking derivatives of $L(\theta)$ with respect to θ, we have

$$\begin{aligned}
\nabla_\theta L &= \mathbb{E}_{X,y,t_o\sim p(t_o)} \left[\int_{l_{1:t_o}} p(l_{1:t_o}|\theta, X) \frac{\partial \mathcal{L}(y, q(\theta, X, l_{1:t_o}))}{\partial \theta} \right. \\
&\quad \left. + \int_{l_{1:t_o}} \mathcal{L}(y, q(\theta, X, l_{1:t_o})) \frac{\partial p(l_{1:t_o}|\theta, X)}{\partial \theta} \right] \\
&= \mathbb{E}_{X,y,t_o\sim p(t_o)} \int_{l_{1:t_o}} p(l_{1:t_o}|\theta, X) \left[\frac{\partial \mathcal{L}(y, q(\theta, X, l_{1:t_o}))}{\partial \theta} \right. \\
&\quad \left. + \mathcal{L}(y, q(\theta, X, l_{1:t_o})) \frac{\partial \log p(l_{1:t_o}|\theta, X)}{\partial \theta} \right] \\
&= \mathbb{E}_{X,y,t_o\sim p(t_o)} \int_{l_{1:t_o}} \int_{l_{t_o+1:T}} p(l_{t_o+1:T}|\theta, X, l_{1:t_o}) \\
&\quad p(l_{1:t_o}|\theta, X) \left[\frac{\partial \mathcal{L}(y, q(\theta, X, l_{1:t_o}))}{\partial \theta} \right. \\
&\quad \left. + \mathcal{L}(y, q(\theta, X, l_{1:t_o})) \frac{\partial \log p(l_{1:t_o}|\theta, X)}{\partial \theta} \right] \\
&= \mathbb{E}_{X,y,t_o\sim p(t_o)} \int_{l_{1:T}} p(l_{1:T}|\theta, X) \left[\frac{\partial \mathcal{L}(y, q(\theta, X, l_{1:t_o}))}{\partial \theta} \right. \\
&\quad \left. + \mathcal{L}(y, q(\theta, X, l_{1:t_o})) \frac{\partial \log p(l_{1:t_o}|\theta, X)}{\partial \theta} \right],
\end{aligned} \tag{9}$$

where T is the maximum possible value of t_o. Since t_o and $l_{1:t_o}$ are mutually independent random variables, we have

$$\begin{aligned}
\nabla_\theta L &= \mathbb{E}_{X,y,l_{1:T}} \left[\mathbb{E}_{t_o\sim p(t_o)} \frac{\partial \mathcal{L}(y, q(\theta, X, l_{1:t_o}))}{\partial \theta} \right. \\
&\quad \left. + \mathbb{E}_{t_o\sim p(t_o)} \mathcal{L}(y, q(\theta, X, l_{1:t_o})) \frac{\partial \log p(l_{1:t_o}|\theta, X)}{\partial \theta} \right].
\end{aligned} \tag{10}$$

Moreover, note that $\log p(l_{1:t_o}|\theta, X)$ can be factorized as:

$$\begin{aligned}
\log p(l_{1:t_o}|\theta, X) &= \log p(l_1|\theta, X) + \log p(l_2|\theta, X, l_1) \\
&\quad + \cdots + \log p(l_{t_o}|\theta, X, l_{1:t_o-1}),
\end{aligned} \tag{11}$$

which can be considered as solving the state distribution over a Markov chain. Then, we have:

$$\begin{aligned}
&\mathbb{E}_{t_o\sim p(t_o)} \mathcal{L}(y, q(\theta, X, l_{1:t_o})) \frac{\partial \log p(l_{1:t_o}|\theta, X)}{\partial \theta} \\
&= \sum_{t'=1}^{T} \left[P(t_o = t')\mathcal{L}(y, q(\theta, X, l_{1:t'})) \sum_{t=1}^{t'} \frac{\partial \log p(l_t|\theta, X, l_{1:(t-1)})}{\partial \theta} \right] \\
&= \sum_{t=1}^{T} \left[\left(\textstyle\sum_{t'=t}^{T} P(t_o = t')\mathcal{L}(y, q(\theta, X, l_{1:t'})) \right) \frac{\partial \log p(l_t|\theta, X, l_{1:(t-1)})}{\partial \theta} \right]
\end{aligned} \tag{12}$$

Furthermore, combining equations (10) and (12), we finally obtain

$$\nabla_\theta L = \underbrace{\mathbb{E}_{X,y,l_{1:T}} \sum_{t=1}^{T} P(t_o = t) \frac{\partial \mathcal{L}(y, q(\theta, X, l_{1:t}))}{\partial \theta}}_{\text{Representation learning objective}, \nabla_\theta L_{\text{rep}}} + \underbrace{\mathbb{E}_{X,y,l_{1:T}} \sum_{t=1}^{T} \left[\left(\sum_{t'=t}^{T} P(t_o = t') \mathcal{L}(y, q(\theta, X, l_{1:t'})) \right) \frac{\partial \log p(l_t | \theta, X, l_{1:(t-1)})}{\partial \theta} \right]}_{\text{Self-rewarding reinforcement learning objective}, \nabla_\theta L_{\text{rl}}}, \tag{13}$$

which proves Theorem 1.

In equation (8), $\nabla_\theta L_{\text{rep}}$ is a standard form of representation learning, namely minimizing the task loss over the features extracted from $l_1, \ldots, l_t$ by the model. In addition, $\nabla_\theta L_{\text{rl}}$ boiling down to a form of policy gradients in reinforcement learning[93], where $p(l_t|\theta, X, l_{1:(t-1)})$ is the action distribution, $r_{t'}$ is the reward received at each time step and $\sum_{t'=t}^{T} r_{t'}$ is the cumulative reward following the execution of an action l_t. Since $r_{t'}$ is defined using the negative values of task loss of the model itself, we name L_{rl} as the self-rewarding reinforcement learning objective.

In conclusion, Theorem 1 reveals that when considering minimizing the expected loss of AdaptiveNN over a vision task, an integration of representation learning and self-rewarding reinforcement learning objectives naturally emerges. The former trains the model to extract deep representations from input visual fixations, while the latter guides the model to strategically select fixation locations within the complex visual environment to minimize the loss. Notably, both of them only leverage the standard task loss, without relying on specialized task formats or additional annotations.

Specific learning algorithm. Given Theorem 1, $\nabla_\theta L_{\text{rep}}$ can be directly used as the gradient signals for learning feature-extraction modules. For the policy gradients $\nabla_\theta L_{\text{rl}}$, as reinforcement learning problems are usually more challenging to solve, we propose an augmented version of its basic formulation. First, we introduce a predefined discount factor $\gamma \in [0, 1]$ (refs. 93–95) and a differential form of rewards, aiming to achieve a flexible modelling of balancing long-term and short-term returns, as well as to stabilize the training process. Thus, on top of equations (5) and (8), the policy gradient rule for updating the model can be expressed as

$$\begin{aligned} \nabla_\theta L_{\text{rl}} &= -\mathbb{E}_{X,y,l_{1:T}} \sum_{t=1}^{T} \left[\left(\sum_{t'=t}^{T} \gamma^{t'-t} (r_{t'} - r_{t'-1}) \right) \nabla_\theta \log p_\pi(l_t | s_{t-1}) \right], \\ r_{t'} &= -P(t_o = t') \mathcal{L}(y, q(\theta, X, l_{1:t'})), \end{aligned} \tag{14}$$

where we have

$$\begin{aligned} \lim_{\gamma \to 0} \sum_{t'=t}^{T} \gamma^{t'-t} (r_{t'} - r_{t'-1}) &= r_t - r_{t-1}, \\ \lim_{\gamma \to 1} \sum_{t'=t}^{T} \gamma^{t'-t} (r_{t'} - r_{t'-1}) &= r_T - r_{t-1}. \end{aligned} \tag{15}$$

On top of equation (14), we actually have

$$\begin{aligned} &\mathbb{E}_{l_t} r_{t-1} \nabla_\theta \log p_\pi(l_t | s_{t-1}) \\ &= r_{t-1} \int_{l_t} p_\pi(l_t | s_{t-1}) \frac{1}{p_\pi(l_t | s_{t-1})} \nabla_\theta p_\pi(l_t | s_{t-1}) \\ &= r_{t-1} \frac{\partial \int_{l_t} p_\pi(l_t | s_{t-1})}{\partial \theta} = 0. \end{aligned} \tag{16}$$

Combining equations (15) and (16), we obtain

$$\begin{aligned} \lim_{\gamma \to 0} \nabla_\theta L_{\text{rl}} &= -\mathbb{E}_{X,y,l_{1:T}} \sum_{t=1}^{T} r_t \nabla_\theta \log p_\pi(l_t | s_{t-1}), \\ \lim_{\gamma \to 1} \nabla_\theta L_{\text{rl}} &= -\mathbb{E}_{X,y,l_{1:T}} \sum_{t=1}^{T} r_T \nabla_\theta \log p_\pi(l_t | s_{t-1}). \end{aligned} \tag{17}$$

When $\gamma \to 0$, the strategy for selecting the next visual fixation tends to be fully short-sighted and is only optimized to maximize the immediate reward r_t. Conversely, $0 < \gamma < 1$ tends to encourage perception strategies that maximally attain the goal within a limited number of fixations. When $\gamma = 1$, AdaptiveNN only focuses on maximizing the final reward r_T, corresponding to the scenarios where abundant resources or energy are available, while the perception process can leverage as many visual fixations as possible to accomplish the task.

Moreover, we introduce a value network V^π to offer a baseline function for reinforcement learning[95,96], which can effectively stabilize training by reducing gradient estimation variance[94,97]. The learning objective of V^π is to predict the expected gains of further observing at each step:

$$\text{Minimize}_{V^\pi} \quad \mathbb{E}\left[V^\pi(s_{t-1}) - \sum_{t'=t}^{T} \gamma^{t'-t} (r_{t'} - r_{t'-1}) \right]^2. \tag{18}$$

Besides, with this goal, $V^\pi(s_{t-1})$ provides a reasonable proxy measure for adaptive termination, as stated in the section 'Architecture of AdaptiveNN'. For example, a relatively small $V^\pi(s_{t-1})$ indicates that even if the model processes more visual fixations, the loss $\mathcal{L}$ measuring the performance of the given task will not show notable further reduction. Hence, it is reasonable to consider concluding observation at that time.

More implementation details on the algorithms and hyperparameters for training AdaptiveNN can be found in Supplementary Sections 4.4 and 4.5.

Studies and discussions on AdaptiveNN's design choices

In pursuit of a comprehensive understanding of our work, we establish a series of evaluations uncovering that the components of AdaptiveNN function as we expect, and that our design markedly outperforms alternative choices.

Strategies for selecting visual fixations. Extended Data Fig. 4a and Supplementary Tables 34 and 41 examine the effectiveness of a broad array of possible strategies for localizing visual fixations. The reinforcement learning algorithm of AdaptiveNN achieves significantly higher validation accuracies than the most competitive baseline across all the scenarios (all $P < 0.0001$), especially with limited numbers of fixations. Although GardCAM has been widely used as a feasible algorithm to visualize the regions relevant to the decision-making of deep networks[34,98], its application in selecting visual fixations does not yield competitive performance against AdaptiveNN, even though it is augmented with a Gaussian mixture model and additional computation. In Supplementary Section 5.5, we further present an in-depth discussion on the reasons for this phenomenon, as well as the inherent limitations of such methods, which sample fixation locations on the basis of fixed goal-directed importance maps (for example, obtained by GardCAM or other algorithms). Moreover, other possible methods for training the fixation selection policy, such as spatial transformer net and Gumbel-Softmax, do not exhibit the potential to approach reinforcement learning. They fail to secure noteworthy gains over predefined non-adaptive policies such as random or Gaussian sampling.

Interpretability. Supplementary Section 5.1 provides a further in-depth examination of AdaptiveNN's interpretability, presenting both qualitative analyses and representative case studies that illustrate how different fixation sequences (for example, from AdaptiveNN and other methods) influence task performance and final success.

AdaptiveNN versus non-adaptive models with down-sampled inputs. In Supplementary Section 5.4, we provide focused comparisons between AdaptiveNN and non-adaptive models with down-sampled inputs, on top of the same backbones. We demonstrate

that down-sampling inputs, although effectively reducing the computational cost, leads to markedly suboptimal efficiency, especially for processing high-resolution, relatively complex images from real-world scenarios.

State values learned by AdaptiveNN. In Extended Data Fig. 4b and Supplementary Tables 42 and 49, we show that the state values predicted by the vision agent of AdaptiveNN are strongly correlated with the test loss of the validation data. This correlation indicates that, for a given test sample whose label is unknown, we can leverage its associated state values as reliable proxies of how far the outputs of our model are from the accurate prediction. This phenomenon is highly consistent with our goal of introducing the value network (Fig. 2d). In this sense, Extended Data Fig. 4c and Supplementary Tables 50 and 53 provide further evidence supporting that the strategy of concluding the observation processes of the samples exhibiting smaller rather than larger state values is beneficial for a higher overall computational efficiency. This strategy underscores the efficacy of the value network in guiding the allocation of computational resources towards optimizing model performance.

Training stability and data efficiency. In Supplementary Section 5.2, we provide additional results elaborating on the training dynamics, hyper-parameter sensitivity and data efficiency of AdaptiveNN. In Supplementary Section 5.3, we further report ablation study results that investigate the impact of AdaptiveNN's different design choices of reward and loss function components.

Comprehensive comparisons with previous methods. Extended Data Fig. 4d,e present system-level comparisons against representative state-of-the-art methods for enhancing the computational efficiency of deep networks. Extended Data Fig. 4d focuses on the recently proposed algorithms that leverage the spatial redundancy or sample-wise redundancy of visual data, whereas Extended Data Fig. 4e considers existing multi-exit models (using the same backbones as AdaptiveNN) characterized by an online-adjustable computational cost. AdaptiveNN outperforms all of them by marked margins when consuming less or comparable amounts of computation, even though the major motivation of our work is to emulate the visual perception behaviours of humans to drive the transition from 'passive' to 'active and adaptive' vision models, instead of attaining optimal engineering performance.

Evaluation tasks for AdaptiveNN

Here we describe the 9 different tasks used for evaluating AdaptiveNN, each associated with 1 or more datasets, yielding 17 benchmarks in total. For all tasks, we held out 20% of the training data to perform a hyper-parameter search, and then put this data back into the training set, reporting final results. When involved, we consider the number of floating-point operations (FLOPs) as the measure of computational cost for the inference of a model.

Computer vision tasks. *Large-scale real-world visual understanding: ImageNet*. ImageNet is a large-scale and diverse dataset of high-quality Internet images[25]. Each image is annotated with a label of its category. The categories are organized according to the WordNet hierarchy[99], covering a wide range of common visual content, including objects, buildings, humans, animals, scenes and so on. ImageNet is a very popular benchmark for evaluating deep learning models[23,26–28], and has been instrumental in advancing computer vision and machine learning research. In this article, we adopted the standard training-validation split, with ~1,280,000 images for training, 50,000 images for validation and 1,000-class annotations. Following the common practice, we used validation accuracy as the performance metric.

Fine-grained visual recognition: six benchmarks. Beyond the general-purpose large-scale visual understanding task, we further probe into AdaptiveNN's nuanced visual discriminative capabilities using six fine-grained recognition tasks. These tasks are characterized by the small differences between classes and substantial variations within each class, such as differentiating between visually very close species of birds or pets against highly diversified backgrounds. Accomplishing them necessitates AdaptiveNN to localize and identify minor, task-dependent signals out of an extensive or even overwhelming multitude of irrelevant visual information. Here we describe the six corresponding datasets we used. For all of them, we adopted the standard training-validation split, and use validation accuracy as the performance metric.

(1) Caltech-UCSD birds-200-2011 (CUB-200-2011)[100] is one of the most widely used fine-grained categorization datasets. It consists of 11,788 images of 200 subcategories belonging to birds, 5,994 for training and 5,794 for testing.
(2) North America Birds (NABirds)[101] contains 48,562 annotated photographs of 400 species of commonly observed birds in North America. Each species has more than 100 photographs, including annotations for males, females and juveniles. All the data are divided into 555 visual categories.
(3) Oxford-IIIT Pet[102] is a 37-category pet dataset with ~200 images for each class. The images are highly diversified in scale, pose and lighting.
(4) Stanford Dogs[103] contains 20,580 images of 120 breeds of dogs from around the world. The dataset is divided into 12,000 images for training and 8,580 images for validation.
(5) Stanford Cars[104] contains 16,185 images of 196 classes of cars. The data are divided into 8,144 training images and 8,041 validation images. The categories are typically built at the level of make, model and year.
(6) FGVC-Aircraft[105] is a benchmark that contains 10,200 images of 102 different classes of aircraft, where each class has 100 images. The data are organized in a four-level hierarchy, namely model, variant, family and manufacturer.

Efficient processing of visual data from real driving scenarios: STSD. The ImageNet and fine-grained recognition datasets are standard visual understanding benchmarks collected from the Internet. As a consequence, in general, many images within them have been centred towards the relevant objects or content by human photographers and users. Similar to human visual systems, AdaptiveNN is also applicable to more general visual perception scenarios. For example, it can process non-object-centric, complex images collected in the wild without specified preprocessing. As a representative example, we considered the task of recognizing traffic signs on the STSD[74]. The dataset consists of 960 × 1,280 road-scene images, captured from real moving vehicles, and the task is to recognize the existence and types of speed limit sign. Note that the targets of interest are generally small, diversely distributed and sometimes not clear. In this article, we used two subsets comprising 747 and 648 images for training and validation, respectively. Validation accuracy is used as the performance metric.

Visual search with diversified task demands: localizing arbitrary digits in multi-digit images. To investigate whether AdaptiveNN has the human-like adaptability of customizing visual perception behaviours conditioned on different task demands, we considered a visual search scenario where the categories and number of targets are assumed to be flexibly changed. Specifically, we created a digit localization dataset by generating 224 × 224 images, each randomly populated with 6–10 28 × 28 MNIST digits[48] against a black background without repetition of digits. We established a large-scale dataset with 500,000 images for training and 50,000 images for validation. To define a visual search task, we specified arbitrary numbers and classes of digits, and trained

our model to identify the locations of these specified digits within each input image. This requires a model to not only recognize correct targets, but also accurately localize multiple targets in a single image. To measure the performance of a given model, we randomly defined many visual tasks and obtained the average success rate on the validation set. Notably, one success means retrieving exactly all the digits demanded by a task from an input, while the success rate of a task is defined as the number of successes divided by the number of all samples.

Image processing in medical scenarios: RSNA pneumonia detection. To demonstrate the efficacy of AdaptiveNN in applications where interpretability holds vital importance, a pneumonia detection scenario was considered. We used the Radiological Society of North America (RSNA) pneumonia dataset, which consists of ~30,000 frontal-view chest radiographs[76]. Each image in the dataset is annotated with image-level labels indicating the presence or absence of pneumonia, as well as bounding boxes for pulmonary opacity, which are visual signals for the disease. The annotations are provided by 18 board-certified radiologists from 16 institutions. In this article, we leveraged the image-level labels to train AdaptiveNN, and compared the locations it fixates on with the pulmonary opacity identified by clinicians to assess its interpretability. The dataset was randomly divided into training and validation sets, following a ratio of 85% and 15%, respectively. The model's diagnostic accuracy was quantified through the AUROC on the validation data.

Embodied AI tasks. *CALVIN long-horizon multi-task language control benchmarks*. We adopt CALVIN[78] to construct the benchmarks for validating the performance of our multi-task, language-guided embodied agent. Within CALVIN, the agent is tasked with executing sequences of actions, each consisting of five subtasks defined through natural language instructions. The model's effectiveness is measured by its average successful length across 1,000 task sequences, with scores ranging from 0 to 5 on the basis of the number of subtasks completed successfully, as detailed in Extended Data Fig. 2b. The CALVIN dataset is organized into four distinct environmental subsets, labelled A to D, each characterized by unique visual backgrounds and object arrangements. Each of these subsets encompasses approximately 24,000 robot manipulation trajectories accompanied by language annotations. We train our embodied MLLMs on these language-annotated trajectories. To thoroughly evaluate the model's ability to imitate and generalize, we conduct experiments under two scenarios: (1) D → D: training and testing within the same environment, and (2) ABCD → D: training on data from all four environments while testing on a single target domain.

Comparisons with human visual perception behaviours. To demonstrate the potential of AdaptiveNN as a valuable tool for investigating human visual cognition, we evaluated humans and AdaptiveNN side by side in the same tests of visual perception behaviours. Specifically, these tests were designed under two goals, namely (1) spatial-wise, examining the locations of visual regions that a human or model fixates on; and (2) sample-wise, examining the difficulty level that a human or model assesses to accomplish the given task on the basis of each individual visual environment. To attain these goals, we conducted three groups of experiments, as described below.

First, through the lens of spatial-wise adaptiveness, we investigated the consistency of the locations of visual fixations selected by AdaptiveNN and humans. We used the SALICON benchmark[79], which consists of 10,000 training images and 5,000 validation images. Every image is annotated with a map of the centres of human gazing. Each gazing centre is treated as a single point in the map. The maps of gazing centres were collected based on the paid Amazon Mechanical Turk crowdsourcing marketplace, with each image observed by ~60 participants. All participants had normal or corrected-to-normal vision and normal colour vision. The images were presented to each participant in a random order, where each image was presented for 5 s. The participants were instructed to explore the image freely by looking at anywhere they wanted to look, with no further instructions on where they should look in the images. The gazing centre locations were obtained by a 100-Hz resampling and processed by excluding the fast-moving data corresponding to saccade processes.

To compare humans and AdaptiveNN in terms of spatial-wise adaptive visual perception, a metric named 'normalized human-like score' was defined. For each image, the average density map of the gazing centres of all ~60 observers was used as the ground-truth distribution of the real focal centres of human vision. We let AdaptiveNN select n visual-fixation regions on top of each image, mimicking the process of freely observing the image for a fixed length of time. Then, we obtained the probability that the ground-truth human gazing centres fall into the visual-fixation regions localized by AdaptiveNN, denoted as p_n^{AdaNN}. Similarly, consider sampling visual-fixation regions following the gazing centre distribution of an arbitrary single person (within ~60 observers), or fully uniformly at random, and then taking the expectations (averaged over ~60 observers and sufficient times of random sampling, respectively). As a result, we had the corresponding expected probabilities $\mathbb{E}[p_n^{\text{Single-human}}]$ and $\mathbb{E}[p_n^{\text{Random}}]$, respectively. Built on this, we defined that

$$\text{Normalized human-like score} = \frac{p_n^{\text{AdaNN}} - \mathbb{E}\left[p_n^{\text{Random}}\right]}{\mathbb{E}\left[p_n^{\text{Single-human}}\right] - \mathbb{E}\left[p_n^{\text{Random}}\right]}. \quad (19)$$

Notably, equation (19) being equal to one indicates that the consistency between AdaptiveNN and the average characteristics of the spatial-wise visual-fixation behaviours of people is approximately the same as the level of an average individual human observer. On the other hand, equation (19) being equal to zero provides a baseline of randomly fixating on the visual environments. In our implementation, normalized human-like scores were calculated on mini-batches of data sampled from the dataset to reasonably reflect their values across different sets of visual environments. We adopted $n = 3$ and a batch size of 64. Moreover, we reported the results on top of the two splits of SALICON (split-1/2 corresponds to the train and validation split of SALICON). They were not particularly distinguished since our model had never been trained on SALICON.

Second, through the lens of sample-wise adaptiveness, we investigated whether our model is consistent with humans in judging which visual environments are relatively easier or more difficult for a given task, and should have less or more attention being paid to observing them. To achieve this, we started by measuring the judgements of difficulty level from humans. Specifically, 10 volunteers (aged between 18 years and 35 years) participated in our experiment (we verified that further increasing the number of participants does not notably affect our findings). All of them had normal or corrected-to-normal vision and normal colour vision. Our studies were approved by the THU S&T Ethics Committee (AI), protocol THU-03-2024-0006, and obtained informed consent. We selected six representative categories of images from the ImageNet validation set. The participants were instructed to assign a 0–100 score to each image according to the difficulty level of the visual recognition task built on this image, where smaller scores mean easier. The order of different categories and the order of images within each category were both randomized for each participant. Each image was presented to a participant for 5 s, after which a corresponding difficulty score was recorded. There was a practice session before formal trials for the participants to get familiar with our experimental procedure, which was identical to the formal trials in all configurations but the scores were not recorded. After the experiment, the scores of each category were normalized on a per-participant basis and averaged across participants. This human-assessed difficulty level was compared with the normalized state values predicted by AdaptiveNN, which reflect our model's judgements on the difficulty level of each visual environment.

Third, we further developed several visual Turing tests[22], leveraging the straightforward human judgements to compare the visual perception behaviours of AdaptiveNN with those of humans. In these tests, real human judges tried to identify the machine, given paired examples of human and machine behaviours. Driven by the previous discussions, our visual Turing tests probed into both the spatial-wise visual-fixation behaviours and the sample-wise visual difficulty-assessment behaviours of our model. For the former, we took the ground-truth density map of human gazing centres for each image from SALICON, and sampled a sequence of three visual-fixation regions, as human behaviours against the machine behaviours of the three fixation locations selected by AdaptiveNN. For the latter, the normalized and averaged human-assessed difficulty scores acquired as aforementioned, and the normalized state values predicted by AdaptiveNN, were rescaled to [0, 100] on a per-class basis, as human and machine behaviours, respectively. In each trial, a human judge was given two paired groups of images (three in each) in a random order, one group comprising human behaviours and the other comprising machine behaviours. The human judge was informed to identify 'which group of images reflects the visual perception behaviours of a machine'. See Supplementary Figs. 12 and 13 for the representative examples of our trials.

The full procedure of 'visual Turing tests' is detailed in Extended Data Fig. 3a. For each visual Turing test concerning the spatial-wise or sample-wise adaptive visual perception behaviours, we considered three types of trial: (1) human versus machine, as described above; (2) human versus human and (3) human versus random. For each trial of (2) and (3), we replaced the group of images corresponding to 'machine' with samples depicting the behaviours of human vision or randomly generated behaviours, yet the participant was still told to distinguish between human versus machine. We established 36 trials for each of (1)–(3), yielding totally 108 trials for each of the two visual Turing tests. These 108 trials were shuffled for every participant, such that (2) and (3) provided randomized control groups as baselines for comparison, and also offered information to validate whether our experimental setups were reasonable. Thirty-nine volunteers (aged between 18 years and 40 years), with normal or corrected-to-normal vision and normal colour vision, participated in our experiment. Our studies were approved by the THU S&T Ethics Committee (AI), protocol THU-03-2024-0006, and we obtained informed consent. We verified that further increasing the number of participants does not notably affect our findings. There was a practice session before real trials. After all trials, each accuracy of (1)–(3) was calculated per participant and aggregated across participants. Notably, 50% accuracy indicates that the sort of behaviours is indistinguishable from those of humans (perfectly human-like), while 100% suggests the inverse case.

Beyond the results presented in the main text, as supplementary results, we also use the widely used MIT1003 (ref. 106) dataset to conduct an additional visual Turing test for model evaluation. MIT1003 was collected using the ETL 400 ISCAN eye tracker at a sampling rate of 240 Hz, involving 15 observers and comprising 1,003 natural indoor and outdoor scenes. The corresponding results are reported in Supplementary Section 5.6, which are consistent with our main findings and further support the human-like visual perception behaviours of AdaptiveNN.

Moreover, in addition to examining whether AdaptiveNN's fixation locations and difficulty-assessment behaviours align with the human visual system, Supplementary Section 5.7 further compares the fixation order of AdaptiveNN with that of humans.

Data availability

Most data used in this study are publicly available, including from ImageNet[25] at https://www.image-net.org/, CUB-200-2011[100] at https://www.vision.caltech.edu/datasets/cub_200_2011/, NABirds[101] at https://dl.allaboutbirds.org/nabirds, Oxford-IIIT Pet[102] at https://www.robots.ox.ac.uk/~vgg/data/pets/, Stanford Dogs[103] at https://paperswithcode.com/dataset/stanford-dogs, Stanford Cars[104] at https://paperswithcode.com/dataset/stanford-cars, FGVC-Aircraft[105] at https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/, STSD[74] at https://www.cvl.isy.liu.se/research/datasets/traffic-signs-dataset/, MNIST[48] at https://paperswithcode.com/dataset/mnist, RSNA pneumonia detection[76] at https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018, CALVIN[78] at https://github.com/mees/calvin, SALICON[79] at http://salicon.net and MIT1003[106] at https://saliency.tuebingen.ai/. A minimum dataset for our visual Turing tests is provided in Supplementary Figs. 12 and 13.

Code availability

Implementation code is available via GitHub at https://github.com/LeapLabTHU/AdaptiveNN (ref. 107).

References

1. Biederman, I. Perceiving real-world scenes. *Science* **177**, 77–80 (1972).
2. Sperling, G. & Melchner, M. J. The attention operating characteristic: examples from visual search. *Science* **202**, 315–318 (1978).
3. Sagi, D. & Julesz, B. 'Where' and 'what' in vision. *Science* **228**, 1217–1219 (1985).
4. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
5. Ölveczky, B. P., Baccus, S. A. & Meister, M. Segregation of object and background motion in the retina. *Nature* **423**, 401–408 (2003).
6. Moore, T. & Armstrong, K. M. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* **421**, 370–373 (2003).
7. Najemnik, J. & Geisler, W. S. Optimal eye movement strategies in visual search. *Nature* **434**, 387–391 (2005).
8. Carrasco, M. Visual attention: the past 25 years. *Vis. Res.* **51**, 1484–1525 (2011).
9. Wolfe, J. M. & Horowitz, T. S. Five factors that guide attention in visual search. *Nat. Hum. Behav.* **1**, 0058 (2017).
10. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. In Proc. *36th International Conference on Neural Information Processing Systems* 23716–23736 (ACM, 2022).
11. *OpenAI Gpt-4 Technical Report* (OpenAI, 2023).
12. *Gemini Team Google Gemini: A Family of Highly Capable Multimodal Models* Technical Report (Google, 2023).
13. Lu, M. Y. et al. A multimodal generative AI copilot for human pathology. *Nature* **634**, 466–473 (2024).
14. Kaufmann, E. et al. Champion-level drone racing using deep reinforcement learning. *Nature* **620**, 982–987 (2023).
15. Zitkovich, B. et al. RT-2: vision-language-action models transfer web knowledge to robotic control. In *Proc. 7th Conference on Robot Learning* (eds Jie, T. & Marc, T.) 2165–2183 (PMLR, 2023).
16. O'Neill, A. et al. Open X-Embodiment: robotic learning datasets and RT-X models: Open X-Embodiment collaboration. In *2024 IEEE International Conference on Robotics and Automation* 6892–6903 (IEEE, 2024).
17. Gehrig, D. & Scaramuzza, D. Low-latency automotive vision with event cameras. *Nature* **629**, 1034–1040 (2024).
18. Chen, A. I., Balter, M. L., Maguire, T. J. & Yarmush, M. L. Deep learning robotic guidance for autonomous vascular access. *Nat. Mach. Intell.* **2**, 104–115 (2020).
19. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
20. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
21. Schäfer, R. et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat. Comput. Sci.* **4**, 495–509 (2024).

22. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
23. Orhan, A. E. & Lake, B. M. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nat. Mach. Intell.* **6**, 271–283 (2024).
24. Vong, W. K., Wang, W., Orhan, A. E. & Lake, B. M. Grounded language acquisition through the eyes and ears of a single child. *Science* **383**, 504–511 (2024).
25. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Computer Vis.* **115**, 211–252 (2015).
26. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds Lourdes, A. et al.) 770–778 (IEEE, 2016).
27. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds Jim, R. et al.) 4700–4708 (IEEE, 2017).
28. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In *International Conference on Learning Representations* (eds Katja, H. et al.) (ICLR, 2021).
29. Dehghani, M. et al. Scaling vision transformers to 22 billion parameters. In *Proc. 40th International Conference on Machine Learning* 7480–7512 (PMLR, 2023).
30. Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. Object detection in 20 years: a survey. *Proc. IEEE* **111**, 257–276 (2023).
31. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds Marina, M. & Tong, Z.) 8748–8763 (PMLR, 2021).
32. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
33. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
34. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
35. Du, Z. et al. ShiDianNao: shifting vision processing closer to the sensor. In *Proc. 42nd Annual International Symposium on Computer Architecture* (ed. David, A.) 92–104 (ACM, 2015).
36. Bai, J., Lian, S., Liu, Z., Wang, K. & Liu, D. Smart guiding glasses for visually impaired people in indoor environment. *IEEE Trans. Consum. Electron.* **63**, 258–266 (2017).
37. Howard, A. G et al. MobileNets: efficient convolutional neural networks for mobile vision applications. Preprint at https://arxiv.org/abs/1704.04861 (2017).
38. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds David, F. et al.) 4510–4520 (IEEE, 2018).
39. Huang, G., Liu, S., Van der Maaten, L. & Weinberger, K. Q. CondenseNet: an efficient DenseNet using learned group convolutions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds David, F. et al.) 2752–2761 (IEEE, 2018).
40. Chen, J. & Ran, X. Deep learning with edge computing: a review. *Proc. IEEE* **107**, 1655–1674 (2019).
41. Wang, X. et al. Convergence of edge computing and deep learning: a comprehensive survey. *IEEE Commun. Surv. Tutor.* **22**, 869–904 (2020).
42. Murshed, M. S. et al. Machine learning at the network edge: a survey. *ACM Comput. Surv.* **54**, 1–37 (2021).
43. Bourzac, K. Fixing AI's energy crisis. *Nature* https://doi.org/10.1038/d41586-024-03408-z (2024).
44. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
45. LeCun, Y. et al. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems* 396–404 (NeurIPS, 1989).
46. Arbib, M. A. *The Handbook of Brain Theory and Neural Networks* (MIT, 1995).
47. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
48. LeCun, Y. *The MNIST Database of Handwritten Digits* (MNIST, 1998); http://yann.lecun.com/exdb/mnist/
49. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
50. Chen, Z. et al. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Zeynep, A. et al.) 24185–24198 (IEEE, 2024).
51. Oquab, M. et al. DINOv2: learning robust visual features without supervision. *Trans. Mach. Learn. Res.* (2024).
52. Ward, D. J. & MacKay, D. J. Fast hands-free writing by gaze direction. *Nature* **418**, 838–838 (2002).
53. Ma, W. J., Navalpakkam, V., Beck, J. M., van den Berg, R. & Pouget, A. Behavior and neural basis of near-optimal visual search. *Nat. Neurosci.* **14**, 783–790 (2011).
54. Henderson, J. M. & Hayes, T. R. Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* **1**, 743–747 (2017).
55. Wolfe, J. M. & Horowitz, T. S. What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* **5**, 495–501 (2004).
56. Hanning, N. M., Fernández, A. & Carrasco, M. Dissociable roles of human frontal eye fields and early visual cortex in presaccadic attention. *Nat. Commun.* **14**, 5381 (2023).
57. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
58. Mnih, V., Heess, N., Graves, A. & Kavukcuoglu, K. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems* 2204–2212 (NeurIPS, 2014).
59. Ba, J., Mnih, V. & Kavukcuoglu, K. Multiple object recognition with visual attention. In *International Conference on Learning Representations* (eds Brian, K. et al.) (ICLR, 2015).
60. Yang, L. et al. Resolution adaptive networks for efficient inference. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Ce, L. et al.) 2369–2378 (IEEE, 2020).
61. Zelinsky, G. J., Chen, Y., Ahn, S. & Adeli, H. Changing perspectives on goal-directed attention control: the past, present, and future of modeling fixations during visual search. *Psychol. Learn. Motiv.* **73**, 231–286 (2020).
62. Wang, Y., Huang, R., Song, S., Huang, Z. & Huang, G. Not all images are worth 16×16 words: dynamic transformers for efficient image recognition. In *Proc. 35th International Conference on Neural Information Processing Systems* 11960–11973 (NeurIPS, 2021).
63. Rao, Y. et al. DynamicViT: efficient vision transformers with dynamic token sparsification. In *35th Conference on Neural Information Processing Systems* 13937–13949 (NeurIPS, 2021).
64. Huang, G. et al. Glance and focus networks for dynamic visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 4605–4621 (2022).
65. Bolya, D. et al. Token merging: your ViT but faster. In *International Conference on Learning Representations* (eds Been, K. et al.) (ICLR, 2023).
66. Gottlieb, J. & Oudeyer, P.-Y. Towards a neuroscience of active sampling and curiosity. *Nat. Rev. Neurosci.* **19**, 758–770 (2018).

67. Navon, D. Forest before trees: the precedence of global features in visual perception. *Cogn. Psychol.* **9**, 353–383 (1977).
68. Chen, L. Topological structure in visual perception. *Science* **218**, 699–700 (1982).
69. Hochstein, S. & Ahissar, M. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* **36**, 791–804 (2002).
70. Ganel, T. & Goodale, M. A. Visual control of action but not perception requires analytical processing of object shape. *Nature* **426**, 664–667 (2003).
71. Oliva, A. & Torralba, A. Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* **155**, 23–36 (2006).
72. Peelen, M. V., Berlot, E. & de Lange, F. P. Predictive processing of scenes and objects. *Nat. Rev. Psychol.* **3**, 13–26 (2024).
73. Touvron, H. et al. Training data-efficient image transformers & distillation through attention. In *Proc. 38th International Conference on Machine Learning* (eds Marina, M. & Tong, Z.) 10347–10357 (PMLR, 2021).
74. Larsson, F. & Felsberg, M. Using Fourier descriptors and spatial models for traffic sign recognition. In *Proc. Image Analysis: 17th Scandinavian Conference, SCIA 2011* (eds Heydn, A. e al.) 238–249 (Springer, 2011).
75. Valliappan, N. et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nat. Commun.* **11**, 4553 (2020).
76. Shih, G. et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artif. Intell.* **1**, 180041 (2019).
77. Li, X. et al. Vision-language foundation models as effective robot imitators. In *International Conference on Learning Representations* (eds Swarat, C. et al.) (ICLR, 2024).
78. Mees, O., Hermann, L., Rosete-Beas, E. & Burgard, W. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robot. Autom. Lett.* **7**, 7327–7334 (2022).
79. Jiang, M., Huang, S., Duan, J. & Zhao, Q. SALICON: Saliency in Context. In *IEEE Conference on Computer Vision and Pattern Recognition* (eds Kristen G. et al.) 1072–1080 (IEEE, 2015).
80. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**, 194–203 (2001).
81. Henderson, J. M. Human gaze control during real-world scene perception. *Trends Cogn. Sci.* **7**, 498–504 (2003).
82. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
83. Kellman, P. J. & Spelke, E. S. Perception of partly occluded objects in infancy. *Cogn. Psychol.* **15**, 483–524 (1983).
84. Spelke, E. S., Breinlinger, K., Macomber, J. & Jacobson, K. Origins of knowledge. *Psychol. Rev.* **99**, 605–632 (1992).
85. Spelke, E. Initial knowledge: six suggestions. *Cognition* **50**, 431–445 (1994).
86. Viola Macchi, C., Turati, C. & Simion, F. Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychol. Sci.* **15**, 379–383 (2004).
87. Simion, F., Di Giorgio, E., Leo, I. & Bardi, L. The processing of social stimuli in early infancy: from faces to biological motion perception. *Prog. Brain Res.* **189**, 173–193 (2011).
88. Ullman, S., Harari, D. & Dorfman, N. From simple innate biases to complex visual concepts. *Proc. Natl Acad. Sci. USA* **109**, 18215–18220 (2012).
89. Stahl, A. E. & Feigenson, L. Observing the unexpected enhances infants' learning and exploration. *Science* **348**, 91–94 (2015).
90. Reynolds, G. D. & Roth, K. C. The development of attentional biases for faces in infancy: a developmental systems perspective. *Front. Psychol.* **9**, 315789 (2018).
91. Bambach, S., Crandall, D., Smith, L. & Yu, C. Toddler-inspired visual object learning. In *Proc. 32nd International Conference on Neural Information Processing Systems* 1209–1218 (ACM, 2018).
92. Orhan, E., Gupta, V. & Lake, B. M. Self-supervised learning through the eyes of a child. In *34th Conference on Neural Information Processing Systems* 9960–9971 (NeurIPS, 2020).
93. Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations* (eds Hugo, L. et al.) (ICLR, 2016).
94. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
95. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. Preprint at https://arxiv.org/abs/1707.06347 (2017).
96. Sutton, R. S., McAllester, D., Singh, S. & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Proc. 13th International Conference on Neural Information Processing Systems* 1057–1063 (ACM, 1999).
97. Silver, D. et al. Mastering the game of go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
98. Wang, L. et al. Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *Nat. Mach. Intell.* **5**, 1356–1368 (2023).
99. Miller, G. A. WordNet: a lexical database for English. *Commun. ACM* **38**, 39–41 (1995).
100. Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. *The Caltech-UCSD Birds-200-2011 Dataset* (Caltech, 2011).
101. Van Horn, G. et al. Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition* (eds Kristen, G. et al.) 595–604 (IEEE, 2015).
102. Parkhi, O. M., Vedaldi, A., Zisserman, A. & Jawahar, C. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (eds Serge, B. et al.) 3498–3505 (IEEE, 2012).
103. Khosla, A., Jayadevaprakash, N., Yao, B. & Li, F.-F. Novel dataset for fine-grained image categorization: Stanford Dogs. In *Proc. CVPR Workshop on Fine-grained Visual Categorization (FGVC)* 2 (2011).
104. Krause, J., Stark, M., Deng, J. & Fei-Fei, L. 3D object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops* (eds Kyros, K. et al.) 554–561 (IEEE, 2013).
105. Maji, S., Rahtu, E., Kannala, J., Blaschko, M. & Vedaldi, A. Fine-grained visual classification of aircraft. Preprint at https://arxiv.org/abs/1306.5151 (2013).
106. Judd, T., Ehinger, K., Durand, F. & Torralba, A. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision* (eds Roberto, C. et al.) 2106–2113 (IEEE, 2009).
107. Yue, Y. LeapLab: LeapLabTHU/AdaptiveNN: official release. *Zenodo* https://doi.org/10.5281/zenodo.16810996 (2025).
108. Caesar, H. et al. nuScenes: a multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11618–11628 (IEEE, 2020).
109. NIH chest X-ray dataset. *Kaggle* www.kaggle.com/datasets/nih-chest-xrays/data (2017).
110. Indian diabetic retinopathy image dataset (IDRiD). *Kaggle* www.kaggle.com/datasets/mohamedabdalkader/indian-diabetic-retinopathy-image-dataset-idrid (2018).
111. COCO 2017 dataset. *Kaggle* www.kaggle.com/datasets/awsaf49/coco-2017-dataset (2017).
112. CUB2002011 dataset. *Kaggle* www.kaggle.com/datasets/wenewone/cub2002011 (2011).

113. ImageNet-1k-valid dataset. Kaggle www.kaggle.com/datasets/sautkin/imagenet1kvalid (2015).
114. The Oxford-IIIT pet dataset. *Kaggle* www.kaggle.com/datasets/tanlikesmath/the-oxfordiiit-pet-dataset (2012).
115. Stanford cars (folder, crop, segment) dataset. *Kaggle* www.kaggle.com/datasets/senemanu/stanfordcarsfcs (2013).
116. FGVC aircraft dataset. *Kaggle* www.kaggle.com/datasets/seryouxblaster764/fgvc-aircraft (2013).
117. Awadalla, A. et al. OpenFlamingo: an open-source framework for training large autoregressive vision-language models. Preprint at https://arxiv.org/abs/2308.01390 (2023).

Acknowledgements

G.H. is supported by the National Key R&D Program of China under grant no. 2024YFB4708200, the National Natural Science Foundation of China under grant nos. U24B20173 and 62276150, and the Scientific Research Innovation Capability Support Project for Young Faculty under grant no. ZYGXQNJSKYCXNLZCXM-I20. S.S. is supported by the National Natural Science Foundation of China under grant no. 42327901. We thank S. Zhang, M. Yao and Y. Wu for helpful discussions and comments on an earlier version of this paper.

Author contributions

G.H. and S.S. initiated and supervised the project. Y.W., Y.Y. (https://orcid.org/0009-0002-0437-7238), Y.Y. (https://orcid.org/0009-0005-3155-1336) and G.H. contributed to the conception and design of the work. Y.W., Y.Y. (https://orcid.org/0009-0002-0437-7238), Y.Y. (https://orcid.org/0009-0005-3155-1336), H.W., H.J., Y.H. and Z.N. contributed to the technical implementation. Y.W., Y.Y. (https://orcid.org/0009-0002-0437-7238), Y.P., M.S., R.L. and Q.Y. contributed to the data acquisition and organization. Y.W., Y.Y. (https://orcid.org/0009-0002-0437-7238), Y.Y. (https://orcid.org/0009-0005-3155-1336), H.W., A.Z. and Z.X. analysed the results. All authors contributed to drafting and revising the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-025-01130-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-01130-7.

Correspondence and requests for materials should be addressed to Shiji Song or Gao Huang.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

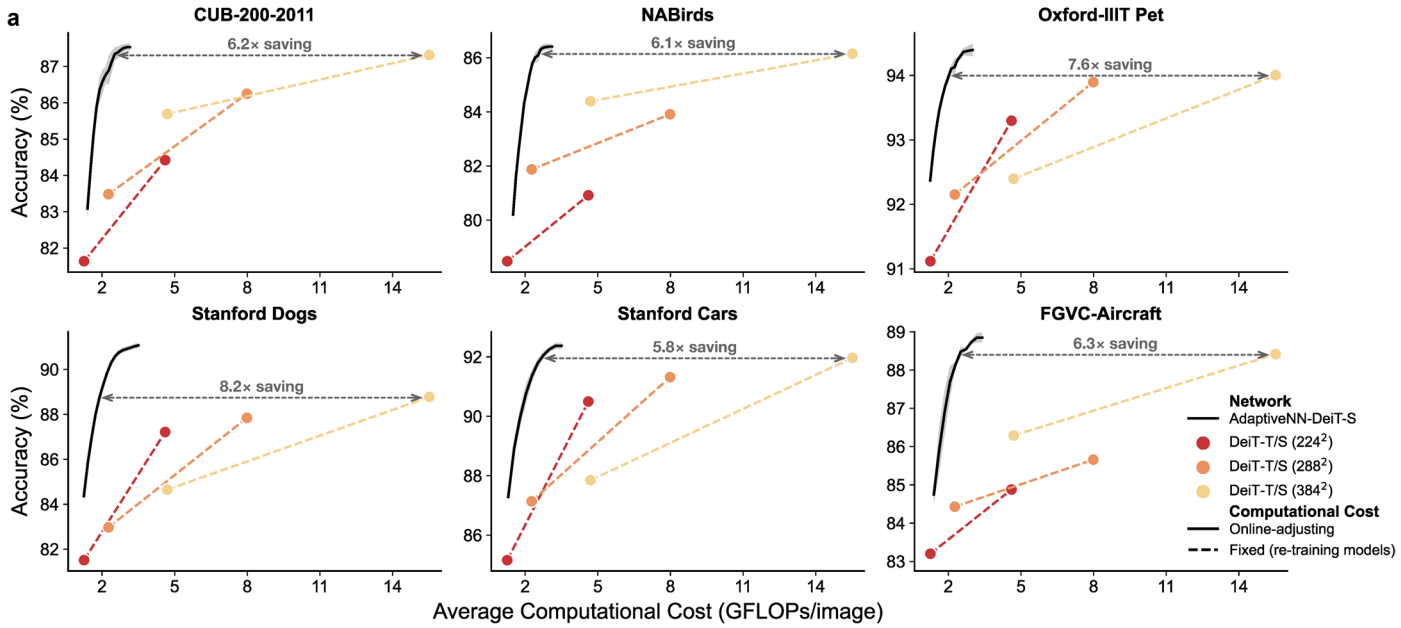
Reprints and permissions information is available at www.nature.com/reprints.

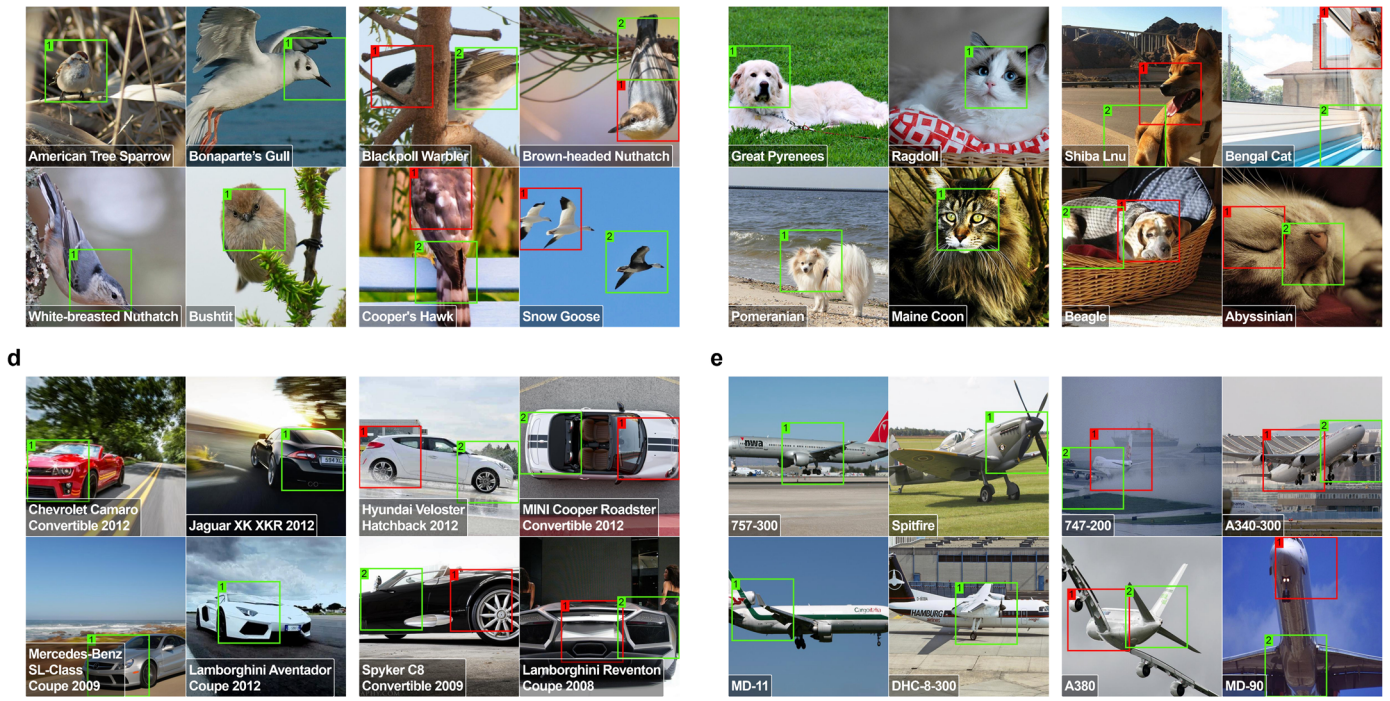
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yulin Wang[1,2], Yang Yue[1,2], Yang Yue[1,2], Huanqian Wang[1], Haojun Jiang[1], Yizeng Han[1], Zanlin Ni[1], Yifan Pu[1], Minglei Shi[1], Rui Lu[1], Qisen Yang[1], Andrew Zhao[1], Zhuofan Xia[1], Shiji Song[1] ✉ & Gao Huang[1] ✉

[1]Department of Automation, Tsinghua University, Beijing, China. [2]These authors contributed equally: Yulin Wang, Yang Yue, Yang Yue.
✉e-mail: shijis@mail.tsinghua.edu.cn; gaohuang@tsinghua.edu.cn

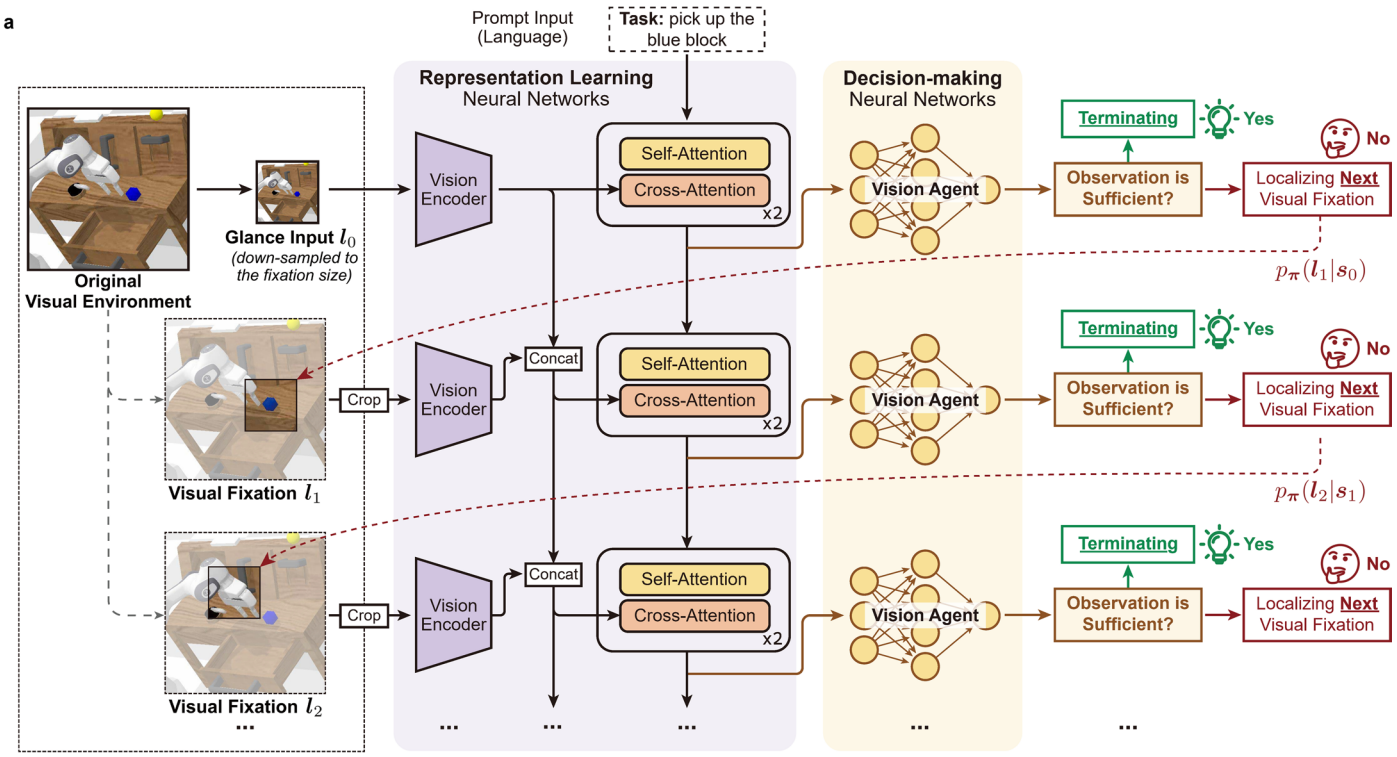
https://doi.org/10.1038/s42256-025-01130-7

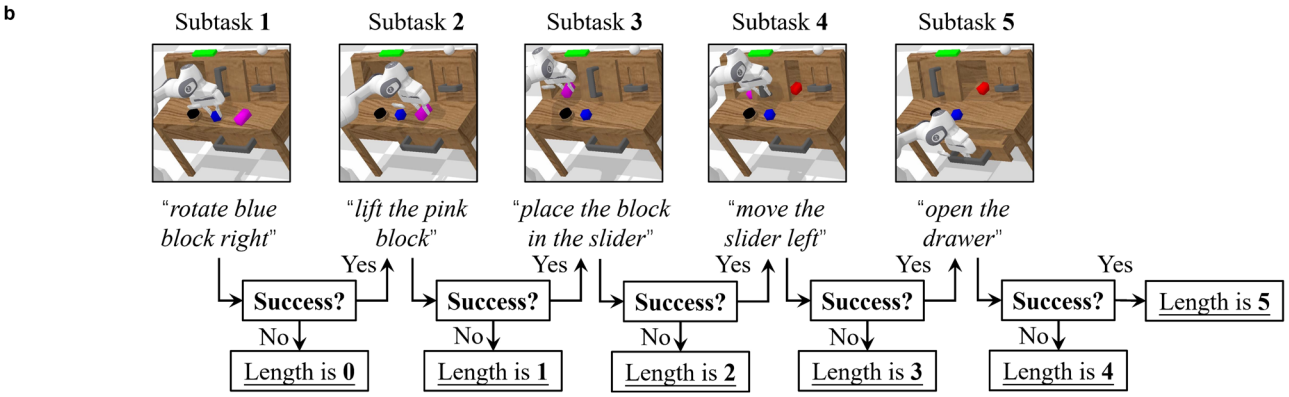




Extended Data Fig. 1 | Results on six fine-grained visual recognition benchmarks. a, Quantitative comparisons of AdaptiveNN and conventional non-adaptive models: Top-1 validation accuracy versus average computational cost for inferring the model. Datasets: CUB-200-2011 [100], NABirds [101], Oxford-IIIT Pet [102], Stanford Dogs [103], Stanford Cars [104], FGVC-Aircraft [105]. The results show means ± standard deviations from five independent trials with different random seeds. Non-adaptive models with varying costs are obtained by modifying model sizes and input resolutions. Here we set the maximum fixation number to be two, which is generally sufficient to accomplish the recognition tasks. **b**–**e**, Qualitative evaluation of the visual fixations chosen by AdaptiveNN-DeiT-S across four datasets: CUB-200-2011, Oxford-IIIT Pet, Stanford Cars, and FGVC-Aircraft. The visualizations adhere to the setups established in Fig. 3a. Images adapted from ref. 112 under a Creative Commons license CC0 1.0; ref. 114 under a Creative Commons license CC BY-SA 4.0; ref. 115 under a Creative Commons license CC0 1.0; refs. 105,116.

https://doi.org/10.1038/s42256-025-01130-7

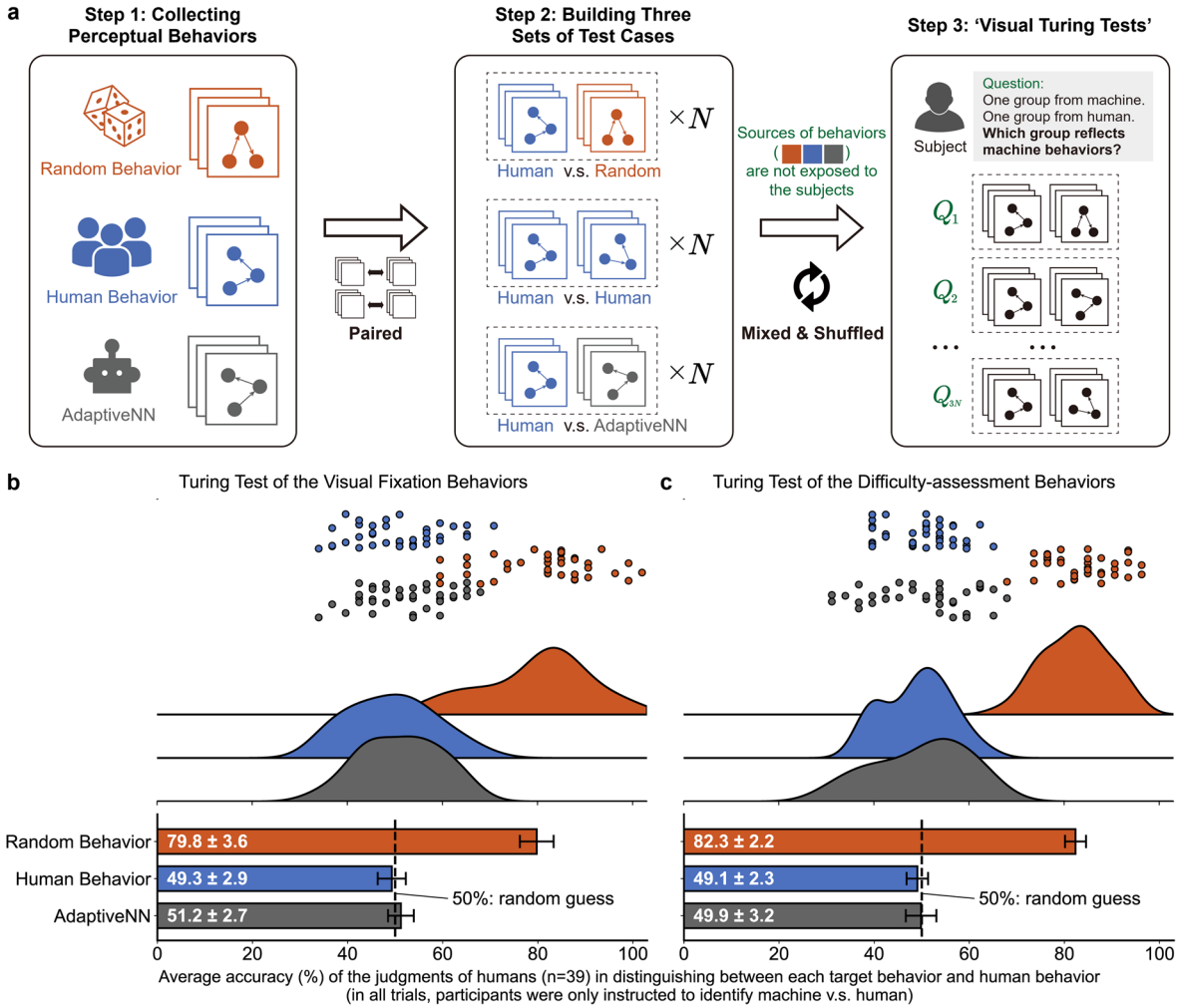




(Average successful length is the average of successful lengths across 1000 task chains)

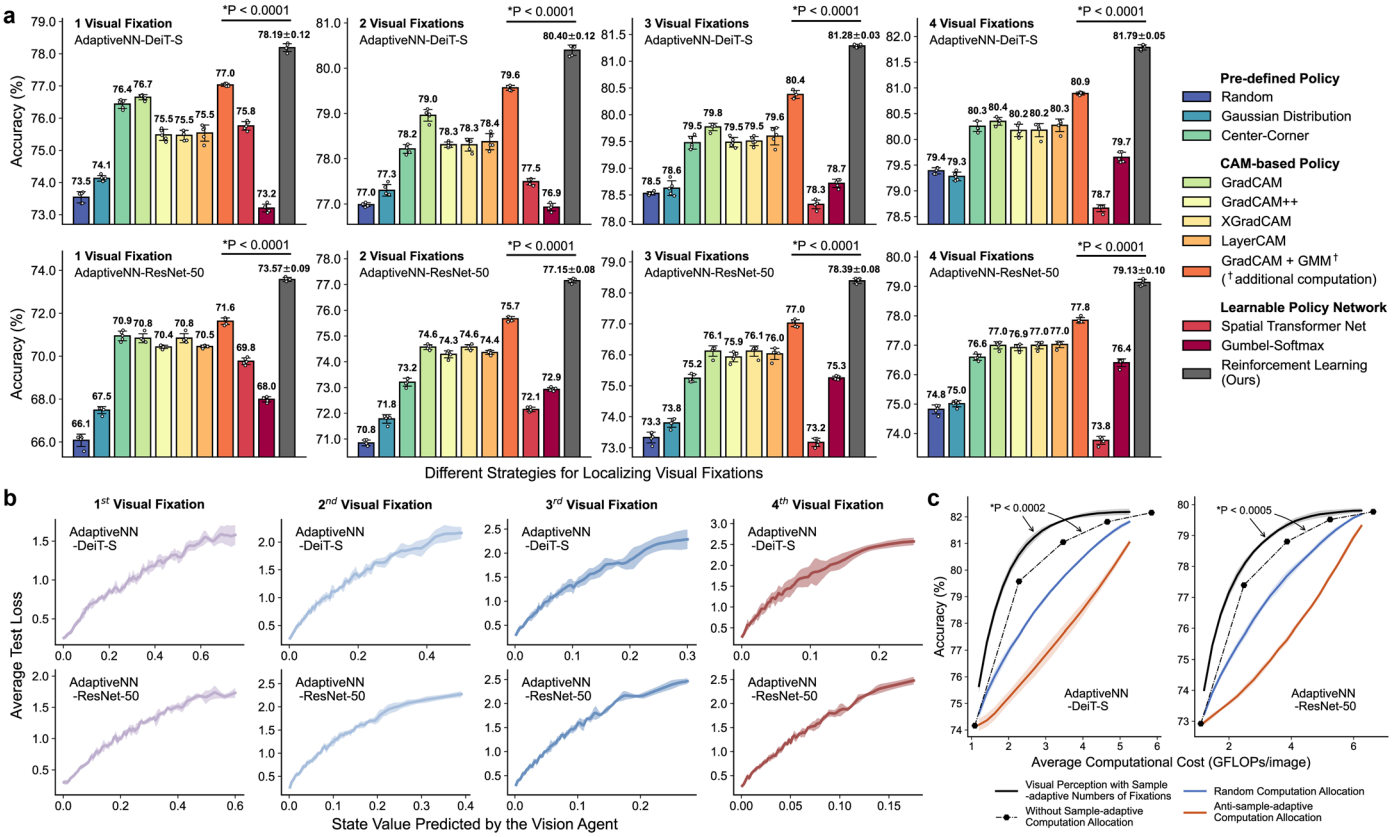
Extended Data Fig. 2 | Details of the experiments based on embodied multimodal large language models (MLLM). a, The network architecture and inference procedure of the AdaptiveNN-based embodied MLLM, which mainly follow RoboFlamingo [77]. The backbone network is based on a pre-trained OpenFLamingo 3B [117]. Each two adjacent network blocks coupled with the shared vision encoder are employed as the perception net of AdaptiveNN. **b**, The metric employed in our experiments on CALVIN. The model performance is quantified as the average successful length (0 to 5) across 1000 5-task sequences. Images are constructed based on the code of ref. 78 (MIT License).

https://doi.org/10.1038/s42256-025-01130-7



Extended Data Fig. 3 | Details of 'visual Turing tests'. a, The full procedure of 'visual Turing tests'. We first collect the visual perception behaviors from real humans, machine (AdaptiveNN), and random generation. Then, we construct multiple trials, each including paired examples of perceptual behaviors. We consider three types of trials: i) human v.s. machine; ii) human v.s. human; and iii) human v.s. random, each corresponding to N trials (we use N=36), yielding totally $3N$ trials for each 'visual Turing test'. Finally, these $3N$ trials are mixed and shuffled for every human judge (n=39). The participants are only instructed to identify the machine behaviors within each trial (for all i)-iii)). Each accuracy of i)-iii) is calculated per participant and aggregated across participants. As a result, i) offers the Turing test results, while ii) and iii) provide randomized control groups as baselines and also validate whether our experimental setups are reasonable. **b**,**c**, Results of visual Turing tests: visual fixation behaviours (**b**) and difficulty-assessment behaviours (**c**). Each data point represents the average identification accuracy of a human judge. Bars show the mean accuracy across human judges and the corresponding 95% confidence interval. Ideal performance is 50%, where the machine is indistinguishable from human behaviors in these binary choice tasks. Data points and their distributions (Gaussian kernel density estimation) are given above the bars.

https://doi.org/10.1038/s42256-025-01130-7



d

Method	Backbone	Average computational cost (GFLOPs/image) ↓	Accuracy (%) ↑
Glance-and-Focus	ResNet-50	2.0	75.6
AdaptiveNN (ours)	ResNet-50	2.0	**77.2**
Confidence interval			**P < 0.0001*
IA-RED²	DeiT-S	3.3	79.1
DynamicViT	DeiT-S	2.9	79.3
Evo-ViT	DeiT-S	2.9	79.4
ATS	DeiT-S	2.9	79.7
A-ViT	DeiT-S	2.9	78.6
Token Merging	DeiT-S	2.7	79.4
Dynamic Perceiver	Multi-Scale DeiT + ResNet-50 + Cross-Attention	2.9	79.9
AdaptiveNN (ours)	DeiT-S	2.9	**81.5**
Confidence interval			**P < 0.0001*

e

Method	Backbone	Accuracy under different computational costs (%) ↑			
		1.25 GFLOPs	1.50 GFLOPs	1.75 GFLOPs	2.00 GFLOPs
MSDNet	DenseNet	72.7	73.7	74.0	74.2
IMTA-MSDNet	DenseNet	73.4	74.3	74.6	74.9
RANet	DenseNet	73.6	74.6	75.0	75.1
Glance-and-Focus	ResNet-50	74.1	75.0	75.3	75.6
AdaptiveNN (ours)	ResNet-50	**74.3**	**75.6**	**76.5**	**77.2**
Confidence interval		**P < 0.0005*		**P < 0.0001*	
DVT	DeiT-S	73.7	75.2	76.8	77.9
AdaptiveNN (ours)	DeiT-S	**76.0**	**77.8**	**79.0**	**79.9**
Confidence interval		**P < 0.0001*			

**Confidence interval indicates the significance at which our method outperforms the strongest baseline.*

Extended Data Fig. 4 | Investigation and ablation studies of the design principles of AdaptiveNN. All the results are reported on ImageNet. See Supplementary Section 3 for the details of comparative baselines. **a**, Efficacy of different methodologies for establishing the fixation localization strategy within AdaptiveNN. For a clean comparison, we train a classifier using only the features from visual fixations, and assume all samples use the same number of fixations, such that the resulting validation accuracy serves as a well-controlled measure to assess the effectiveness of each variant. Moreover, we consider an extensive variety of baselines for comparison, including selecting fixations using i) pre-defined rules; ii) goal-directed importance maps obtained by CAM (class activation map) algorithms; iii) CAM augmented with a Gaussian mixture model (GMM); and iv) policy networks learned using other algorithms. **b**, Average test loss corresponding to the validation data with different state values predicted by the *Vision Agent* in AdaptiveNN. We examine the state values taken from every step of sequential perception processes. **c**, Comparisons of different termination criteria for concluding the sequential perception process of AdaptiveNN. The term 'anti-' refers to the inverse of our proposed method (detailed in Supplementary Section 4.1), namely terminating the observation process for samples with relatively higher state values. Exact P (> 0.0001) values: 0.00018, 0.00047. **d**–**e**, Comparisons with representative methodologies designed to improve deep learning models' computational efficiency. Specifically, **d** evaluates against baselines that leverage the spatial redundancy or sample-wise redundancy in visual data. **e** examines models with multi-exit architectures (using the same backbones as AdaptiveNN) that allow for online computational cost adjustments. Exact P (> 0.0001) value in **(e)**: 0.00034. In **a**–**c**, the results show means ± standard deviations from five independent trials with different random seeds. *Two-sided independent samples *t*-test.