SCIENCE CHINA
Technological Sciences

• Article •

https://doi.org/unknown-doi

Ability Decomposition and Difficulty Quantification of Visual Tasks: Towards Systematic Evaluations of Artificial General Intelligence

Shaoyang Cui[1], Xinyi He[4,5], Jiaheng Han[4,5], Zhenliang Zhang[5] & Yujia Peng[*2,3,5]

[1] *Yuanpei College, Peking University, Beijing* 100871, *China;*
[2] *School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing* 100871, *China;*
[3] *Institute for Artificial Intelligence, Peking University, Beijing* 100871, *China;*
[4] *School of Intelligence Science and Technology, Peking University, Beijing* 100871, *China;*
[5] *State Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence, Beijing* 100080, *China*

Received January 11, 2016; accepted April 6, 2016; published online January 1, 2016

With the rapid development of multi-modal foundation models and the pursuit of Artificial General Intelligence (AGI), there is a growing need for corresponding evaluation systems. Systematic AGI evaluation requires tasks that encompass a wide range of ability dimensions and difficulty levels. However, although many benchmarks exist, the field still lacks a quantification system to assess ability decompositions or difficulty levels. Here, we took the visual domain as a starting point and proposed an explainable system for Task Ability Decomposition and Difficulty Level quantification of Vision (TADDL-V). Using large language models, TADDL-V decomposed the visual abilities required for a given task and leveraged statistical data to map between ability sets and task difficulty levels. The estimated ability masses align with human intuition, and TADDL-Vs task difficulty estimates are empirically validated against aggregated human comparisons of task difficulty. Furthermore, we proposed an AGI visual evaluation task set, AGI-V70, comprising 70 composite visual tasks that incorporate visual abilities across a broad spectrum of task difficulties. Together, TADDL-V serves as a prototype for ability decomposition and task difficulty level quantification, which are essential for future AGI evaluations.

Ability decomposition, Task difficulty quantification, AI Evaluation, Artificial General Intelligence

Citation: Cui S Y, He X Y, Han J H, Zhang Z L, Peng Y J. Ability Decomposition and Difficulty Quantification of Visual Tasks: Towards Systematic Evaluations of Artificial General Intelligence. Sci China Tech Sci, 2025, –, https://doi.org/unknown-doi

0cm

1 Introduction

Recent advances in large-scale and multi-modal foundation models have significantly reshaped the landscape of artificial intelligence (AI). With the emergence of the Transformer architecture [1], the discovery of scaling laws [2], increasing computational resources, and the widespread application of reinforcement learning from human feedback (RLHF) [3], models like GPTs have evolved from achieving successes in isolated domains to the foundation of a rapidly diversifying ecosystem [4]. New generations of reasoning models, such as OpenAI o1, o3, o4-mini, and DeepSeek-R1, have demonstrated increasingly general-purpose reasoning [5,6].

Rapid advancement has also intensified interest in achieving Artificial General Intelligence (AGI). The human-level performance of GPTs in various text-based tasks has led to the argument that the era of AGI is imminent [7].

However, several studies have highlighted existing limi-

*Corresponding author. Email: yujia_peng@pku.edu.cn

tations of GPT models in specific domains. For example, even the most advanced GPT-4 model struggles with complex tasks (e.g., [8]). Moreover, while GPTs can sometimes provide correct answers as assistive systems, they often produce explanations that are vague, overly verbose, or irrelevant. Similarly, GPT models appear to be "right for the wrong reasons," indicating a possible lack of robust reasoning ability [9]. From the perspective of social intelligence, the entire series of GPT models may still fall significantly short of human-level performance [10]. Consequently, the debate over whether current foundation models have yet reached the level of general intelligence underscores the need for a scientifically sound and rational approach to evaluating the level of intelligence of an agent to determine whether it possesses "general intelligence." A quantitative and systematic approach would be a key not only to define AGI, but also to guide the steps of AGI development [11].

Importantly, systematic evaluations of AGI or even current foundation models require benchmarks that cover a broad range of difficulties and ability dimensions. Classic AI evaluations followed a task-oriented paradigm [12–16], which involves assessing the agents abilities through specific sets of tasks. However, given the "general" nature of AGI, no limited task sets can be claimed "complete" in terms of AGI evaluation, and an ability-oriented approach is needed to comprehensively evaluate an AGI model given a finite task set [11, 17, 18]. Moreover, to thoroughly and comprehensively evaluate an AGI, it is crucial to establish a clear correspondence between testing tasks and the agent's abilities, along with scientifically quantifying the difficulty levels [11, 12, 16, 19, 20].

Although numerous benchmarks exist, the field still lacks a quantification system to assess the difficulty level or ability decompositions. During the past few decades, influential studies on various computer vision topics have repeatedly invoked the notions of high-level versus low-level visual tasks and of hard versus easy problems, underscoring the communitys need for well-defined, difficulty-calibrated, and ability-aware metrics [21–28].

For instance, in video action recognition, the extraction of semantic video information was considered as a high-level task, while treating regional features as low-level [21]. Other studies regarded object concepts such as faces as high-level, and edges or blobs as low-level [25, 26]. Conversely, in visual commonsense reasoning, conventional recognition and categorization of objects, scenes, actions, and events were relegated to lower-level status, whereas fluency, causality, and physical reasoning were deemed high-level [27]. Multimodal understanding research had also labeled task variants as hard or easy [28], yet these annotations were based on ad hoc judgments and binary distinctions without theoretical grounding or quantitative validation. The existing discrepancies, sometimes outright contradictions, reveal that high- and low-level categories of tasks are far from uniform and that task difficulty cannot be reduced to a simple dichotomy. A unified theoretical framework is therefore required to systematically stratify visual capabilities and to quantify task difficulty.

In addition to task difficulty quantification, ability decomposition is also essential for comprehensive AGI evaluations. With the same task description, the underlying abilities involved may be drastically different. For example, with a task of "finding a banana", it could refer to locating a banana in a picture that tests object recognition alone, or referring to retrieving a banana from a refrigerator after being given a verbal instruction that engages visual, linguistic, physical, and commonsense reasoning skills simultaneously (**Figure 1**). Existing AI benchmarks were typically constructed within specific domains such as vision or language, where the range of task types and underlying abilities was relatively constrained. It has often been unnecessary to explicitly specify the cognitive or perceptual abilities required by each task, as the tasks collectively focus on a limited, well-understood skill set. As a result, benchmarks like ImageNet [29] or AG News [30] evaluate performance without decomposing tasks into constituent abilities.

Overall, the lack of difficulty quantification and ability decomposition becomes a major limitation in AGI evaluation, making it challenging to interpret agent performance, to compare results across tasks, or to pinpoint areas of weakness.

Here, we propose that a unified theoretical framework is fundamental to enable meaningful and generalizable AGI evaluation, allowing for the systematic stratification of visual capabilities and the quantification of task difficulty. Each task should be described in two dimensions: **the required abilities** involved, and a **quantified difficulty level**. These annotations provide interpretability in AI evaluations, allowing for fair comparisons and making it possible to diagnose an AIs strengths and weaknesses across a broad ability spectrum.

Specifically, we took the visual domain as a starting point and proposed an explainable system for Task Ability Decomposition and Difficulty Level quantification of Vision (TADDL-V). Utilizing large language models (LLMs), TADDL-V decomposed the visual abilities required for a given task and leveraged statistical data to map between ability sets and task difficulty levels. The estimated ability masses aligned with human intuition, and TADDL-Vs task-difficulty estimates were empirically validated against aggregated human judgments. Furthermore, we released a benchmark of 70 composite visual tasks with annotated abilities

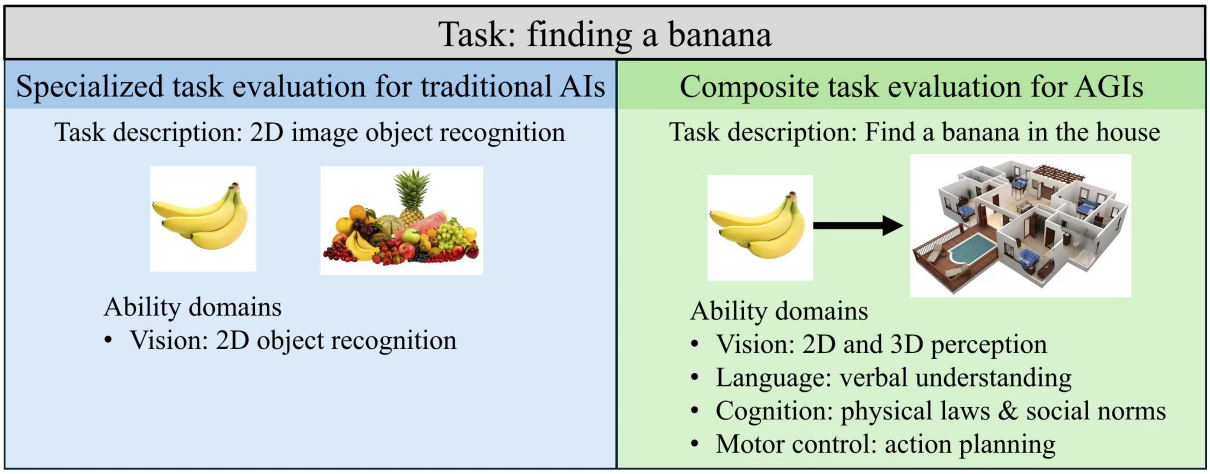

Figure 1 An illustration of the contrast between specialized and composite tasks. A visual task of finding a banana can be framed as a specialized task (e.g., image classification) or as a composite task that involves interpreting verbal instructions, reasoning about object locations, and physically interacting with the environment.

and difficulty values, **AGI-V70**, covering a wide range of real-world vision reasoning scenarios. Our work effectively revealed the difficulty distribution of the selected benchmarks, providing reference and data support for the construction and optimization of future benchmarks. Together, the current work provides a scalable and interpretable foundation for broader AGI benchmarking efforts, presenting a prototype for generalizable systems that can model the abilitydifficulty space across diverse domains and task types.

2 Methods

2.1 Task, Abilities, and Task Difficulty

In this subsection, we clarify three key concepts used throughout this paper: **Task**, **Ability**, and **Task Difficulty**. These definitions provide the conceptual foundation for the proposed TADDL-V framework.

Task. Unlike traditional benchmarks composed of narrow, domain-specific tasks (e.g., digit classification or sentiment analysis), AGI-oriented benchmarks should capture complex, real-world scenarios. Thus, AGI evaluations are inherently based on *composite tasks*, often requiring the coordination of multiple perceptual, cognitive, and motor abilities.

To represent each task, we define a structured triplet:

1. **Task Name:** A brief description of the task.
2. **Initial State (Input):** The initial state of the environment.
3. **Target State (Output):** The desired state after successful completion of the task.

Ability. Defining and categorizing abilities is inherently more nuanced. In real-world scenarios, most tasks are composite and cannot be completed using a single atomic skill. Taking the task "Find a banana in the house" (**Figure 1**) as an example, an embodied general agent must invoke multiple visual abilities: object recognition (detecting the banana in a scene), spatial reasoning (understanding its location in 3D space), and path planning (navigating to the object based on 2D visual inputs). Each of these high-level abilities can be further decomposed. For instance, object recognition may involve conceptual understanding, color recognition, shape perception, and edge detection.

Since the way abilities are decomposed can vary considerably across annotators and applicationsdepending on factors such as granularity, domain knowledge, or task framingwe adopted a more pragmatic strategy. Specifically, we defined a fixed, human-designed **Ability Set**, providing a standardized reference for decomposing tasks into abilities. By anchoring our analysis to a predefined set, we can perform ability reasoning and difficulty estimation in a way that is both consistent across annotators and reproducible across experiments.

Task Difficulty. Quantifying task difficulty is another key challenge, as it is inherently relative to the agent being evaluated. A task that is trivial for a human may be highly complex for a machine, and vice versa. To establish a consistent evaluation reference, we defined task difficulty in this paper from a *human-centric perspective*. Specifically, the perceived difficulty of each task was grounded in human judgments, reflecting our broader commitment to human-aligned AGI evaluation, as humans remain the most general and capable intelligent agents currently known.

While difficulty is often treated as a fuzzy or subjective concept, we aim to develop a quantitative, scalable represen-

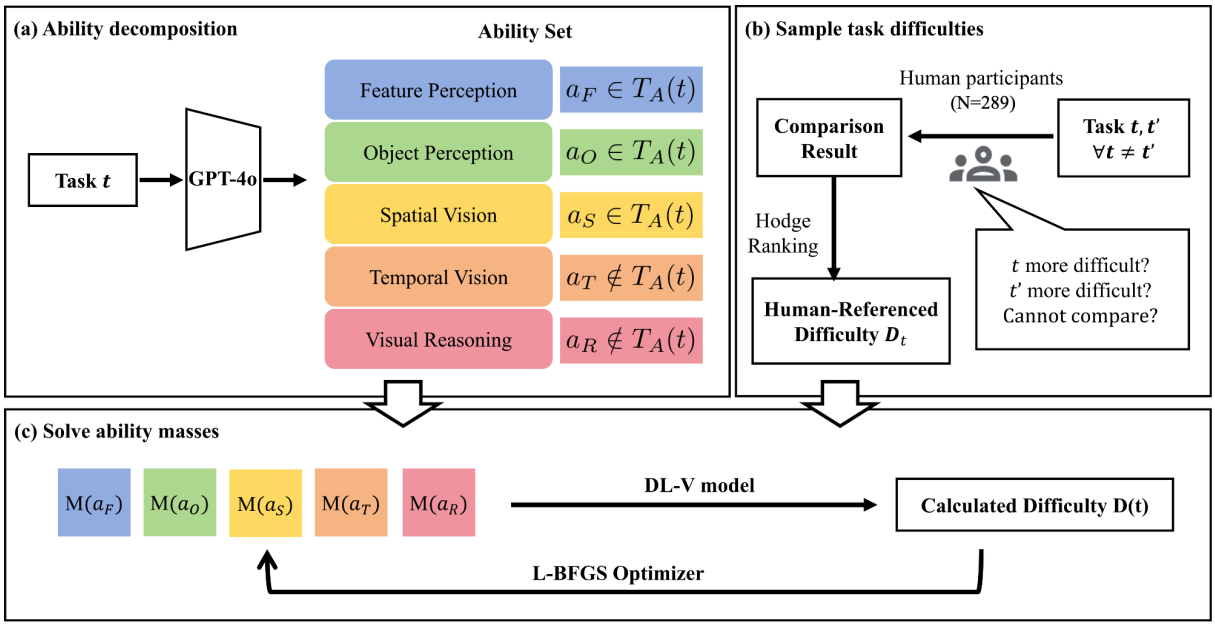

Figure 2 An illustration of the TADDL-V system pipeline. (a) Ability decomposition process: Utilizing GPT-4o to decompose the required abilities for each task t_i in a (given) task set. (b) Conducting a survey and processing the pairwise comparison results with the Hodge rank algorithm to obtain the global rank (difficulty level s). (c) Using optimization methods to solve Eq. 5 for appropriate $F(A)$ values.

tation. In the remainder of this paper, we introduce methods to approximate the difficulty of each task using pairwise human comparisons and then anchor these estimates to a model grounded in required abilities.

2.2 Extract task-relevant abilities

In this work, we collected human annotations to identify the abilities required to solve each task. Specifically, we conducted a structured survey in which human annotators were asked to evaluate the vision-related abilities required for a given task. In parallel, we also employed a pretrained large language model (GPT-4o) to perform the same annotation task, allowing us to compare human and model judgments (see **Appendix** for detailed ability decomposition protocol). To ensure generality, we experimented with different levels of granularity when defining the ability set. Details of these variants can be found in **Appendix**.

For each task, we aggregated human annotations using a simple majority rule: if more than 50% of annotators marked a particular ability as required, we considered that ability to be part of the human-assigned ability set for that task. In subsequent figures and analyses, we refer to this aggregated result as **Human**. Formally, we define the mapping from task t to its required abilities as:

$$T_A(t) = \{a_i \mid \text{solving task } t \text{ requires ability } a_i\}. \tag{1}$$

2.3 Associating task difficulty with required abilities

Building on our previous design and intuitive considerations, we propose that the difficulty of a task can be understood as a function of the ability mass it involves. In other words, the required set of abilities collectively determines an abstract mapping from abilities to a scalar difficulty value, reflecting how demanding the task is. Our central hypothesis is that tasks requiring higher-level or more complex abilities tend to be more difficult. To capture this, we associate each ability a with a scalar value called its **mass**, denoted $M(a)$, which reflects the ability's intrinsic difficulty weight.

In practice, abilities often exhibit semantic and functional overlap, particularly when defined at different levels of granularity. For example, object recognition and feature perception may both be required by a task and yet share significant redundancy (overlapping). Hence, we adopted an inclusionexclusion-style formulation to avoid double-counting overlapping contributions:

$$D(t) = \sum_i M(a_i) - \sum_{i<j} \mathrm{Cor}(a_i, a_j) + \sum_{i<j<k} \mathrm{Cor}(a_i, a_j, a_k) - \cdots + (-1)^{n+1} \mathrm{Cor}(a_1, \ldots, a_n), \tag{2}$$

where

$$\mathrm{Cor}(S) = \sum_{a_i \in S} \log M(a_i), \qquad S \subseteq \{a_1, \ldots, a_n\}. \tag{3}$$

We refer to this formulation as the **DL-V model**, which serves as the core of our task difficulty estimation framework.

Terms $\mathrm{Cor}(a_i, a_j, \ldots)$ represent pairwise or higher-order correlation penalties between overlapping abilities, explicitly correcting for their joint contribution. In our implementation, we estimated these correlation terms using an information-theoretic heuristic derived from ability co-occurrence statistics across tasks. Specifically, if two abilities frequently co-occur across the dataset, we penalize their joint contribution to avoid over-counting.

To apply the DL-V model in practice, we need to address (1) how to assign a numeric difficulty $D(t)$ to each task, as well as (2) how to determine the ability mass terms $M(a)$ and correlation penalties Cor, which we detail in the following subsections.

2.4 Task difficulty quantification

Although many intuitive strategies exist for assessing task difficulty, manually calibrating difficulty scores for all tasks at scale is infeasible. Therefore, we begin by defining a finite, well-curated set of tasks to serve as the basis for difficulty modeling and optimization.

Large language models (LLMs), such as GPT-series models, have demonstrated impressive capabilities in understanding and generating natural language [31]. Prior studies highlight their emergent abilities in task understanding, procedural decomposition, and knowledge-grounded generation [32, 33]. In particular, their strength in controlled text generation makes them ideal tools for synthesizing realistic task descriptions [34]. Inspired by this, recent works such as **TaskBench** [35] have shown the feasibility of automating task generation using LLMs.

Leveraging GPT-4o for task generation and subsequent human filtering and refinement, we constructed a benchmark containing 70 composite, vision-centric daily tasks, which we refer to as **AGI-V70**. Each task in AGI-V70 follows the structured format introduced earlier: task name, initial state, and target state.

To obtain human-perceived difficulty levels for these tasks, we conducted a large-scale survey in which participants were asked to perform pairwise comparisons. As shown in **Figure 2 (b)**, each of approximately 289 participants was presented with 50 randomly sampled task pairs and asked to select the more difficult one. In total, we collected 2,415 pairwise comparisons, ensuring that every task pair was evaluated at least five times by different individuals.

We applied the Hodge ranking algorithm [36] to the aggregated pairwise comparison data to derive a global scalar difficulty score for each task. The resulting ranking score $s^{\mathrm{origin}}(t)$ reflects the relative difficulty of task t across the full set. Higher values indicate greater perceived difficulty. To ensure all estimated ability masses remain positive, we normalize the Hodge scores using the following offset:

$$s(t) \leftarrow s^{\mathrm{origin}}(t) - \min_i s^{\mathrm{origin}}(i) + \epsilon, \tag{4}$$

where ϵ is a small positive constant to avoid zero values. Thus, $s(t)$ serve as natural targets for the DL-V model: $D_t := s(t)$. (See **Figure 2** for an overview.)

2.5 Ability mass quantification

Given the DL-V model defined in Eq. 2, we aim to estimate the mass $M(a)$ for each ability $a \in A$, where A is the predefined ability set. We formulate an optimization problem that fits the model-predicted task difficulties $D(t)$ to the human-perceived difficulty scores D_t, derived from the Hodge ranking over AGI-V70.

$$\begin{aligned} \min_M \quad & \sum_{t=1}^{70} d(D_t, D(t; M)) \\ \text{s.t.} \quad & M(a_i) > 0, \quad \forall a_i \in A. \end{aligned} \tag{5}$$

Here, $d(x_1, x_2)$ is a dissimilarity metric between the target difficulty D_t and the model prediction $D(t)$; we use the L2 norm in our experiments. The constraint $M(a_i) > 0$ is imposed to ensure that all abilities are associated with non-negative intrinsic difficulty weights, consistent with the semantics of the mass concept (The full algorithmic procedure is provided in the **Appendix**).

3 Experiments

3.1 Questionnaire consistency analysis

As described in Section 2.4, we conducted a survey in which participants were asked to compare the difficulty of task pairs. Given the inherent subjectivity in human judgment, we evaluated the internal consistency of the survey responses to ensure data reliability.

To quantify consistency, we defined an internal consistency metric α based on the distribution of responses for each task pair. Each pairwise comparison has three possible outcomes:

1. C_1: participants judged Task A as more difficult.
2. C_2: participants judged Task B as more difficult.
3. C_3: participants judged the two tasks as equally difficult.

A task pair is considered to meet a given consistency level α if the following condition holds:

$$\max_i (C_i + C_3) \geq \alpha \sum_{j=1}^{3} C_j, \quad i \in \{1, 2\}. \tag{6}$$

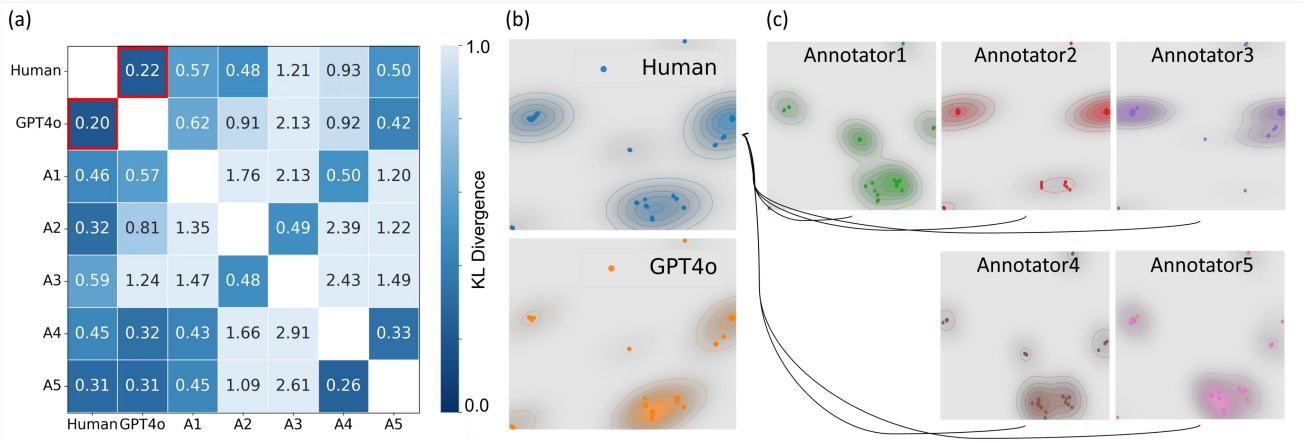

Figure 3 Comparisons of ability annotations from human and GPT-4o. (a) KL-divergence between each distribution and the aggregated human result. (b) Aggregated human annotations (top) vs. GPT-4o annotations (bottom). (c) Distribution of ability annotations from individual human experts. The point clouds are obtained by applying UMAP for dimensionality reduction on the one-hot encoded ability annotations across all tasks in AGI-V70.

Based on the value of α, we categorized consistency into four levels: **strongly consistent** ($\alpha = 1.0$), **consistent** ($\alpha = 0.8$), **weakly consistent** ($\alpha = 0.6$), and **inconsistent** (otherwise). Applying this classification to the *pairwise difficulty comparisons collected during the survey*, we found that over **72%** of task pairs were rated as either strongly consistent or consistent, while only **2.1%** were labeled inconsistent (see Table 1). The high degree of agreement among participants demonstrates the reliability of the collected pairwise difficulty judgments and supports their suitability for downstream modeling.

Table 1 Consistency Distribution of Pairwise Difficulty Judgments from the Survey.

Consistency Level	Proportion
Strongly consistent ($\alpha = 1.0$)	28.3%
Consistent ($\alpha = 0.8$)	43.4%
Weakly consistent ($\alpha = 0.6$)	26.3%
Inconsistent (otherwise)	2.1%

3.2 Ability annotation: human vs. GPT-4o

To facilitate experimental focus, we restricted our analysis to the visual domain. We defined a set of five core visual abilities, referred to as Vision-5 (V5), listed in Table 2. Both human experts and GPT-4o were asked to annotate the abilities required for each of the 70 tasks in the AGI-V70 benchmark.

An interesting observation emerges from **Figure 3 (c)** that individual human experts exhibit significant variability in their annotations. To analyze the individual differences, we compared each annotators distribution of selected abilities to the aggregated consensus from all human experts. We further compared these distributions to that produced by GPT-4o. Using KL-divergence as a distance metric (shown in **Figure 3 (a)**), we found that human experts are highly inconsistent with one another, where the distribution from each expert diverged substantially from the group consensus. In contrast, GPT-4o's annotation closely matches the aggregated human result.

This finding highlights two key insights. First, GPT-4o, as a powerful pretrained statistical model, appears to approximate the mean of human judgment distributions. Moreover, using GPT-4o for ability annotation provides a stable and scalable alternative to mitigate inter-human variability. These results justify our choice to rely on GPT-4o annotations in following experiments.

3.3 Exploration of ability set design

The construction of the ability set is a central element of the TADDL-V framework. The current experiment investigated how the structural design of ability sets influenced the frameworks capacity to approximate sampled task difficulty values (Eq. 2) and, ultimately, to achieve effective optimization in solving Eq. 5.

Specifically, we examined two key dimensions in the design of ability sets: the *breadth* of a set refers to the range of distinct cognitive or perceptual abilities it encompasses, and the *granularity* reflects how fine-grained or decomposed each ability is defined. To systematically explore these dimensions, we constructed two complementary families of ability sets. The **FO2-FO5** family exhibits low breadth, focusing solely on two core visual abilities, Feature Perception and Object Detection, while varying in granularity. The **Vision-2Vision-5** family, by contrast, maintains broader coverage, expanding from the comprehensive Vision-5 base and incrementally refining its components. The structure of these sets is illustrated in **Figure 4 (a)**.

Given the demonstrated reliability of GPT-4o annotations

Table 2 The five visual abilities in the Vision-5 set.

Ability	Explanation
Feature Perception	Identifying and matching low-level features such as shape, color, and texture.
Object Perception	Detecting, segmenting, and recognizing discrete objects in a scene.
Spatial Vision	Understanding spatial relations (e.g., distance, orientation, layout) between objects for navigation and interaction.
Temporal Vision	Perceiving and interpreting visual changes over time, including motion, sequences, and prediction.
Visual Reasoning	Performing logical inference over visual information, including decision making, problem solving, and abstraction.

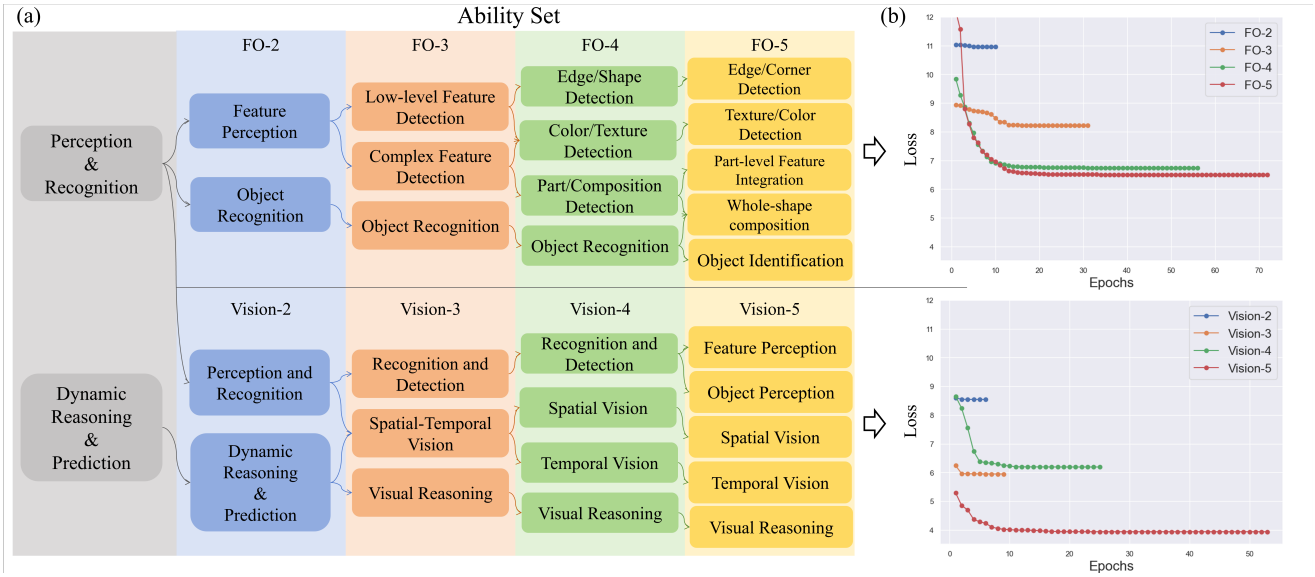

Figure 4 Comparisons between ability sets. (a) Different ability set configurations. From coarse-grained (FO) to fine-grained (Vision); (b) Loss curves for solving Eq 5 under different ability set configurations.

(Section 3.2), we used it to label all 70 tasks in AGI-V70 under each of the FO2-FO5 and Vision-2-Vision-5 configurations and fit the DL-V model using L-BFGS optimization. We evaluated the L2-norm loss between the predicted difficulty values (from Eq. 2) and the sampled difficulty values from Hodge ranking. The results are shown in **Figure 4 (b)**.

Two clear trends emerged from the results:

1. **Higher granularity improves fit**: As the ability sets became more fine-grained (e.g., from Vision-2 to Vision-5), the final optimization loss decreased, indicating better task difficulty modeling. An exception was Vision-4, which converged to a slightly higher loss than Vision-3; however, the overall trend remained consistent.

2. **Greater breadth improves fit**: Comparing **Figure 4 (b)**: The Vision sets, which covered a wider range of abilities, consistently outperformed the FO sets, highlighting the importance of designing ability sets that comprehensively span the task space.

Together, these findings emphasize that both *coverage* and *granularity* of the ability set significantly affect the representational capacity and optimization quality of the TADDL-V framework.

3.4 Effectiveness of the TADDL-V framework

We evaluated the effectiveness of the TADDL-V framework from two complementary perspectives. First, we examined whether the incorporation of correlation terms in Eq. 2 improved the models capacity to estimate task difficulty accurately. Second, we assessed the frameworks generalization ability by testing on unseen tasks and evaluating its consistency with human judgments beyond the training set.

3.4.1 Correlation terms improve performance

To account for the potential overlap between abilities within a chosen ability set, inspired by mutual information in information theory, we introduced correlation terms into the DL-V model (**Eq. 2**. These terms were designed to capture the joint contribution of overlapping abilities, drawing on principles such as inclusionexclusion and entropy decomposition. To validate their effectiveness, we fit the DL-V model under the Vision-5 ability set while varying the maximum order of correlation terms included (from 2nd to 5th order). As shown

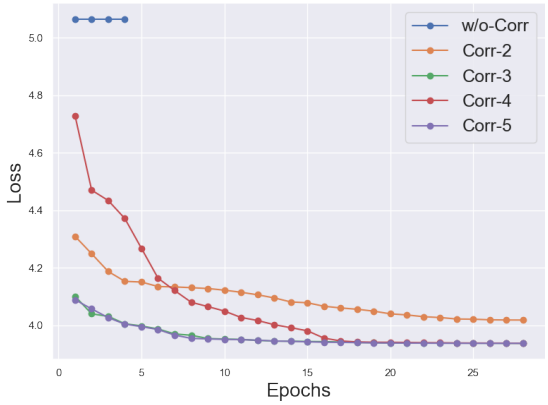

Figure 5 Regression loss of TADDL-V on AGI-V70 with increasing correlation order. Note that, due to early stopping, different groups loss curves have varying lengths.

in **Figure 5**, the regression loss consistently decreased with the inclusion of higher-order terms, confirming that modeling ability overlap improves performance.

3.4.2 Consistency with human ratings on unseen task difficulties

To assess the agreement between modelpredicted task difficulties and human judgments, we employed **alignment rate** to quantify the proportion of task pairs for which the numerical ranking of difficulties (produced by HodgeRank, DLV, or other methods) agreed with the majority of human pairwise comparisons.

Let $\mathcal{T}$ denote the set of tasks and $\mathcal{P} = \{(t_i, t_j) \mid t_i, t_j \in \mathcal{T}, , i < j\}$ the set of task pairs. For each pair (t_i, t_j), we encoded the majority human judgment as $y_{ij} \in -1, 0, +1$:

$$y_{ij} = \begin{cases} +1, & \text{if a strict majority of interviewees judged } t_i \text{ harder than } t_j, \\ -1, & \text{if a strict majority of interviewees judged } t_j \text{ harder than } t_i, \\ 0, & \text{otherwise (no strict majority or judged equally difficult).} \end{cases} \tag{7}$$

Given a difficulty scoring function $D : \mathcal{T} \to \mathbb{R}$ (e.g., HodgeRank, DL–V with human annotations, DL–V with GPT–4o annotations), we defined the calculated order for each pair as

$$\hat{y}_{ij} = \operatorname{sgn}(D(t_i) - D(t_j)), \quad \operatorname{sgn}(0) = 0. \tag{8}$$

The **alignment rate** was defined as the fraction of pairs for which the model–implied order matched the human majority:

$$\text{AlignRate}(D) = \frac{1}{|\mathcal{P}'|} \sum_{(i,j)\in\mathcal{P}'} \mathbf{1}[\hat{y}_{ij} = y_{ij}], \tag{9}$$

where $\mathcal{P}' = \{(i, j) \in \mathcal{P} \mid y_{ij} \in \{-1, +1\}\}$ was the subset with a strict–majority judgment.

We first evaluated alignment rates on the full AGI–V70 dataset with 70 composite visual tasks. Our results show that TADDL–V achieved values comparable to the HodgeRank baseline and close to human annotations, whether using human or GPT–4oderived inputs. Notably, HodgeRank itself reached at most about 75%, which suggests that TADDLV nearly reached the practical performance limit set by variability in human judgments(See the Appendix for the complete results).

Because all 70 tasks in AGI–V70 were also used to solve for the ability masses of the five vision abilities in the DL–V model, these results could have reflected in–sample performance. To assess generalization, we constructed a validation set of 10 previously unseen tasks. Each task was inserted into the global AGI–V70 ranking and was compared against 70 existing tasks by human annotators, producing 10×70 new pairwise task difficulty comparisons.

In both cases, the **alignment rate** was computed as the fraction of aligned pairs among all evaluated pairs-$\binom{70}{2}$ for AGI–V70 and 10×70 for the validation set-yielding a single consistency score.

Results on this validation set showed that alignment rates exceeded 80% when TADDL–V was driven by either human or GPT–4o annotations. These findings demonstrated that TADDL–V not only reproduced human–consistent difficulty rankings on its training set but also generalized robustly to novel tasks, thereby underscoring the reliability of its learned difficulty representation.

Furthermore, we compared TADDL-Vs performance against an artificial neural network (ANN) baseline; experimental results are shown in **Appendix**. Together, the findings

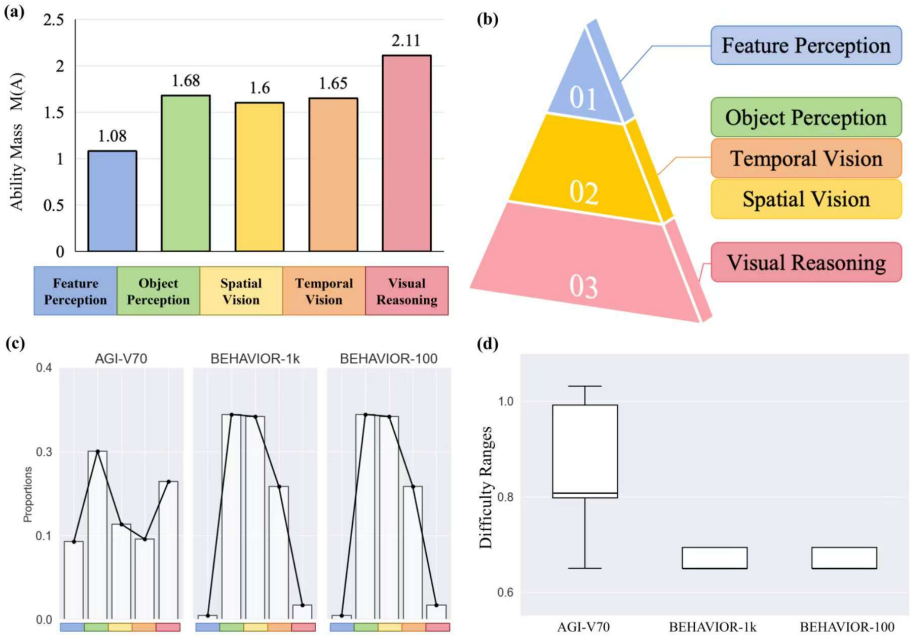

Figure 6 (a) The optimized weighted average masses of the five ability sets, and (b) their alignment with the semantic categories (basic, intermediate, advanced). (c) Distribution of assessed abilities across BEHAVIOR-series and AGI-V70 tasks. (d) Task difficulty ranges estimated by TADDL-V.

demonstrate the rationality, robustness, and effectiveness of the TADDL-V system in quantifying task difficulty.

4 Results

4.1 Explainable results: alignment between solved masses and designed semantics

Within the TADDL-V framework, we employed GPT-4o as the mapping function $T_A(t)$ to establish taskability correspondences in the AGI-V70 set (**Figure 2 (a)**), and further validated these annotations through human expert labeling. Using the L-BFGS optimizer to solve Eq. 5 with the L2-norm as the loss function, we obtained average weighted masses for the five ability sets (**Figure 6 (a)**).

The optimized results revealed a clear and interpretable pattern: the relative magnitudes of the ability masses align well with the semantics of our original ability design. Specifically, Feature Perception consistently emerged with the lowest weight, reflecting its role as a basic perceptual ability; Object Perception, Spatial Vision, and Temporal Vision occupied the middle range, corresponding to intermediate-level perceptual demands; while Visual Reasoning exhibited the highest weight, resonating with its nature as an advanced cognitive ability (**Figure 6 (b)**).

We stress that this outcome does not imply a definitive or exhaustive hierarchy of visual abilities. As noted in Section 3.3, the construction of a complete and stable ability set remains an open challenge. Our present contribution is to demonstrate that, even with a prototypical ability design, the optimization procedure produces ability masses that are meaningfully consistent with human intuition, which provides preliminary evidence that TADDL-V is capable of revealing interpretable structures in the ability space, and it highlights a promising direction for future research: developing more systematic strategies to enumerate and organize ability sets in a principled manner.

4.2 Applying TADDL-V for benchmark ability and difficulty analysis

We applied the TADDL-V system to two representative benchmarks in the BEHAVIOR series-BEHAVIOR-100 [37] and BEHAVIOR-1K [38], together with our AGI-V70 dataset, in order to analyze their visual ability coverage and task difficulty profiles. The BEHAVIOR benchmarks were selected due to their shared design principles: both focus on simple, composite tasks situated in household environments, such as grasping, cleaning, collecting, and placing objects. These benchmarks are representative of low-level embodied task settings that emphasize practical utility but exhibit structural repetitiveness.

As shown in **Figure 6 (c)**, the TADDL-V analysis revealed two key findings. First, the results for BEHAVIOR-100 and BEHAVIOR-1K are highly consistent across both ability distribution and difficulty range, indicating that TADDL-V produces stable and robust estimates under scale variation. This aligns with expectations, as BEHAVIOR-1K is a scaled-

up extension of BEHAVIOR-100, expanding the number of scenes, objects, and tasks without altering the underlying task formats. Second, the BEHAVIOR benchmarks display an imbalanced ability distribution. Abilities such as *Feature Perception* and *Visual Reasoning* are underrepresented, while mid-level abilities like *Object Perception* and *Spatial Vision* dominate. Additionally, the overall difficulty range is narrow, reflecting limited task variance, as many tasks share similar structural templates despite diverse content. As a result, these benchmarks offer limited capacity to differentiate agents across a wide spectrum of cognitive competencies.

By contrast, AGI-V70 spans a broader spectrum of task types and difficulty levels, enabling a more comprehensive evaluation of visual capabilities. These results demonstrate the effectiveness of TADDL-V in characterizing composite-task benchmarks and underscore its value in revealing both the strengths and the limitations of ability coverage and difficulty diversity.

5 Discussions

The current work introduced TADDL-V as a prototype system for explainable ability decomposition and task difficulty quantification in the visual task domain. TADDL-V provides a foundation for structured and meaningful AGI evaluation by introducing a principled mapping from composite visual tasks to an interpretable ability space, and defining task difficulty via optimized ability-weighted metrics. The empirical results demonstrate not only the feasibility of this framework but also the potential to capture human-aligned notions of complexity and competence.

Beyond the empirical validation, this work offers two main contributions to the development of general AI benchmarks and evaluations. First, TADDL-V serves as a foundational tool for systematic, interpretable, and generalizable intelligence benchmarking, addressing a key gap in AGI evaluation. While tasks are traditionally assessed in isolation or without clarity about what abilities they represent, TADDL-V enables fine-grained, explainable annotations of both ability demands and difficulty levels, supporting both top-down (ability-oriented) and bottom-up (task-driven) evaluation strategies, offering a structured pathway for interpreting agent performance in terms of underlying cognitive competence.

Second, applying TADDL-V to a diverse set of composite tasks (AGI-V70) revealed an interpretable gradient of visual abilities, where the optimized ability masses aligned well with the semantic design of our ability set. This pattern resonates with cognitive theories of vision and enables the placement of tasks along an intuitive, easy-to-hard spectrum, which is rarely supported by existing benchmarks. The framework provides both descriptive clarity and diagnostic granularity in evaluating AI systems. Moreover, by applying TADDL-V to representative general-purpose benchmarks with composite visual tasks, we observe that classic benchmarks predominantly emphasize mid-level perceptual abilities and exhibit limited variance in difficulty, while high-level abilities such as visual reasoning or temporal inference remain largely untested. By making these gaps explicit, TADDL-V offers a systematic means to diagnose the limitations of existing benchmarks and to assess their adequacy in evaluating general visual intelligence.

Finally, TADDL-V represents not a fixed solution but a generalizable foundation for future AGI evaluation research. The current implementation focused on five visual abilities, but the same methodology can be extended to incorporate multi-modal abilities such as language, memory, and motor control.

While TADDL–V establishes a structured and interpretable foundation for ability-based task analysis, several limitations remain. First, while the TADDL–V was designed for evaluating composite tasks in embodied environments, it does not currently address domains such as abstract reasoning (e.g., Raven's Progressive Matrices, where embodiment may not be a necessary prerequisite). Extending the framework to encompass such tasks is a vital direction for future work, as it would provide a more complete assessment of visual intelligence. Second, the present implementation focuses primarily on five core visual abilities, leaving multi-modal and interactive dimensions-such as language grounding, memory, and embodied reasoning-outside the current scope. Third, the ability gradient observed in **Section 4.1** is contingent upon the predefined ability set and the assumed correlation structure among abilities. Accordingly, it should be regarded as a relative ordering within the present framework, rather than as definitive evidence of a fixed or universal cognitive hierarchy. Fourth, the construction of the ability set itself remains preliminary. In this work, we only sketched two pragmatic strategies-maximizing breadth and ensuring fine-grained decomposition without providing a stable methodology for enumerating and organizing the entire ability space. Developing a more systematic and potentially hierarchical organization of abilities (e.g., tree-structured decomposition) is an important direction for future work. Finally, the current study focused on the group-level evaluation but did not include individual differences in the scope. Future studies may consider individual differences in the internal representations of the task and, therefore, the solution strategies.

In conclusion, TADDL-V provides a principled step toward interpretable and scalable AGI evaluation, offering a

foundation that future work can further refine and extend.

6 Data Availability

The complete set of tasks in AGI-V70, along with their corresponding difficulty levels, can be found at the Data.

All the experiments discussed in the main paper and the appendix were provided along with their corresponding programs, at the Github for details. Additionally, we've implemented a program based on our TADDL-V system that allows users to analyze and quantify the difficulty of arbitrary tasks.

7 Acknowledgments

The work was supported by the National Science and Technology Major Project (2022ZD0114900), the National Natural Science Foundation of China (32471151,32200854), and the Young Elite Scientists Sponsorship Program (2021QNRC00) to Yujia Peng.

Conflict of interest The authors declare that they have no conflict of interest.

1 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, 2017. 60006010
2 Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv, 2020, 2001.08361. doi:10.48550/arXiv.2001.08361
3 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst, 2022, 35: 2773027744
4 Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv, 2023, 2303.08774. doi:10.48550/arXiv.2303.08774
5 Jaech A, Kalai A, Lerer A, et al. OpenAI o1 system card. arXiv, 2024, 2412.16720. doi:10.48550/arXiv.2412.16720
6 Guo D, Yang D, Zhang H, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv, 2025, 2501.12948. doi:10.48550/arXiv.2501.12948
7 Bubeck S, Chadrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv, 2023, 2303.12712. doi:10.48550/arXiv.2303.12712
8 Collins K M, Jiang A Q, Frieder S, et al. Evaluating language models for mathematics through interactions. Proc Natl Acad Sci USA, 2024, 121: e2318124121
9 Holterman B, van Deemter K. Does ChatGPT have theory of mind?. arXiv, 2023, 2305.14020. doi:10.48550/arXiv.2305.14020
10 Wang J, Zhang C, Li J, et al. Evaluating and modeling social intelligence: a comparative study of human and AI capabilities. arXiv, 2024, 2405.11841. doi:10.48550/arXiv.2405.11841
11 Peng Y, Han J, Zhang Z, et al. The Tong Test: evaluating artificial general intelligence through dynamic embodied physical and social interactions. Eng, 2023. doi:10.1016/j.eng.2023.07.006
12 Wang P. The Evaluation of AGI Systems. In: Proceedings of Artificial General Intelligence, 2010. doi:10.2991/AGI.2010.33
13 Xu B, Ren Q. Artificial open world for evaluating AGI: a conceptual design. In: Proceedings of Artificial General Intelligence, Cham, 2023. 452463
14 Potapov A, Scherbakov O, Bogdanov V, et al. Analyzing elementary school olympiad math tasks as a benchmark for AGI. In: Proceedings of Artificial General Intelligence, Cham, 2020. 279289
15 Lebiere C, Gonzalez C, Warwick W. A comparative approach to understanding general intelligence: predicting cognitive performance in an open-ended dynamic task. In: Proceedings of the 2nd Conference on Artificial General Intelligence, 2009. 711
16 Morris M R, Sohl-Dickstein J, Fiedel N, et al. Position: levels of AGI for operationalizing progress on the path to AGI. In: Proceedings of the Forty-first International Conference on Machine Learning, Vienna, 2024.
17 Bugaj V, Goertzel B. AGI Preschool: a framework for evaluating early-stage human-like AGIs. In: Proceedings of the 2nd Conference on Artificial General Intelligence, 2009. 1217
18 Hernndez-Orallo J. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artif Intell Rev, 2017, 48: 397447
19 Zhong W, Cui R, Guo Y, et al. AGIEval: a human-centric benchmark for evaluating foundation models. In: Proceedings of Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, 2024. 22992314
20 Kwan W C, Zeng X, Wang Y, et al. M4LE: a multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. arXiv, 2023, 2310.19240. doi:10.48550/arXiv.2310.19240
21 Sadanand S, Corso J J. Action bank: a high-level representation of activity in video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 12341241
22 Schulze Buschoff L M, Akata E, Bethge M, et al. Visual cognition in multimodal large language models. Nat Mach Intell, 2025, 7: 96-106
23 Zhao R, Yuan Q, Li J, et al. Sce2DriveX: a generalized MLLM framework for scene-to-drive learning. arXiv, 2025, 2502.14917. doi:10.48550/arXiv.2502.14917
24 Shi J, Nie M, Lin W, et al. A novel Riemannian sparse representation learning network for polarimetric SAR image classification. arXiv, 2025, 2502.15302. doi:10.48550/arXiv.2502.15302
25 Le Q V. Building high-level features using large scale unsupervised learning. In: Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, 2013. 85958598
26 Bertasius G, Shi J, Torresani L. High-for-low and low-for-high: efficient boundary detection from deep object features and its applications to high-level vision. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 504512
27 Zhu Y, Gao T, Fan L, et al. Dark, beyond deep: a paradigm shift to cognitive AI with humanlike common sense. Eng, 2020, 6: 310345
28 Yang Y, Zhang S, Shao W, et al. Dynamic multimodal evaluation with flexible complexity by vision-language bootstrapping. arXiv, 2024, 2410.08695. doi:10.48550/arXiv.2410.08695
29 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 248255
30 Zhang X, Zhao J J, LeCun Y. Character-level convolutional networks for text classification. arXiv, 2015, 1509.01626. doi:10.48550/arXiv.1509.01626
31 Kalyan K S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat Lang Process J, 2024, 6: 100048
32 Zhao W X, Zhou K, Li J, et al. A survey of large language models. arXiv, 2023, 2303.18223. doi:10.48550/arXiv.2303.18223
33 Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. Trans Mach Learn Res, 2022. doi:10.48550/arXiv.2206.07682
34 Sun J, Tian Y, Zhou W, et al. Evaluating large language models on controlled generation tasks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 2023. 31553168

35 Shen Y, Song K, Tan X, et al. TaskBench: benchmarking large language models for task automation. In: Proceedings of ICLR 2024 Workshop on Large Language Model (LLM) Agents, Vienna, 2024. doi:10.48550/arXiv.2311.18760
36 Jiang X, Lim L H, Yao Y, et al. Statistical ranking and combinatorial Hodge theory. Math Program, 2011, 127: 203244
37 Srivastava S, Li C, Lingelbach M, et al. BEHAVIOR: benchmark for everyday household activities in virtual, interactive, and ecological environments. In: Proceedings of Conference on Robot Learning, 2022. 477490
38 Li C, Zhang R, Wong J, et al. BEHAVIOR-1K: a human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. arXiv, 2024. doi:10.48550/arXiv.2403.09227