

北京市地方标准
家庭环境通用人工智能视觉理解能力测试规范

Specification of Visual Understanding Ability Test of Artificial
General Intelligence in Family Environment

点击此处添加与国际标准一致性程度的标识

文稿版次选择

×××× - ×× - ××发布

×××× - ×× - ××实施

目 次

前言	II
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 测试指标体系	2
4.1 测试指标概述	2
4.2 任务泛化能力	2
4.3 自主生成任务	3
5 测试任务维度	4
5.1 概述	4
5.2 客体属性及关系理解	5
5.3 视觉语言联合解译	5
5.4 状态理解	5
5.5 视觉常识理解	5
5.6 事件理解	5
5.7 行为意图理解	6
5.8 情绪理解	6
6 测试流程	6
6.1 概述	6
6.2 测试准备	7
6.3 测试执行	7
6.4 测试结果	7
附录 A （资料性） 家庭环境通用人工智能视觉理解能力测试结果示例	9
A.1 测试任务	9
A.2 测试结果	9
参考文献	10

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由北京通用人工智能研究院、北京大学、电子科技大学、合肥综合性国家科学中心人工智能研究院（安徽省人工智能实验室）提出。

本文件由北京市科委、中关村管委会归口。

本文件由北京市科委、中关村管委会组织实施。

本文件起草单位：

本文件主要起草人：

家庭环境通用人工智能视觉理解能力测试规范

1 范围

本文件规定了适用家庭环境的通用人工智能视觉理解能力的测试规范，描述了测试方法等内容。

本文件适用于指导通用人工智能系统开发方、用户方以及第三方等相关组织对家庭环境中通用人工智能系统的视觉理解能力开展测试工作。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 20000.1—2014 标准化工作指南 第1部分：标准化和相关活动的通用术语

GB/T 41867—2022 信息技术 人工智能 术语

3 术语和定义

GB/T 41867—2022界定的以及下列术语和定义适用于本文件。

3.1

通用人工智能 artificial general intelligence, AGI

具有高效的学习和泛化能力、能够根据所处的复杂动态环境自主产生并完成非事先定义的任务、具备人类价值体系的人工智能。

3.2

家庭环境 home environment

由人类居住空间及其附属设施构成的典型室内场景集合，包含但不限于起居室、卧室、厨房、卫生间等功能区域，涵盖家具、家电、日用品等家居物品，以及家庭成员、宠物等动态要素的有机整体。其空间布局与物体分布需符合现实居住场景的物理逻辑与社会习俗。

注：区别于实验室环境或工业场景，强调非结构化、动态变化的真实生活场景特性。

3.3

视觉理解能力 visual understanding ability

智能系统通过获取家庭环境视觉数据，基于视觉信息，实现对于物体、场景、事件和意图的理解，支持视觉任务的定义和完成。

3.4

任务 task

被调度的训练或推理对象。

注：任务用于完成一个相对独立的业务功能。一个任务属于且仅属于一个作业。

[来源：GB/T25000.23—2019，4.12，有修改]

3.5

模态 modality

信息呈现的类型，包含图像、视频、文本和声音等。

4 测试指标体系

4.1 测试指标概述

家庭环境通用人工智能视觉理解能力测试指标体系包括任务泛化能力指标以及自主生成任务能力指标，每个指标由子指标组成，见图1。在实施评估过程中，应根据不同类型的通用人工智能算法，在不同质量特性下设置具体评估指标。

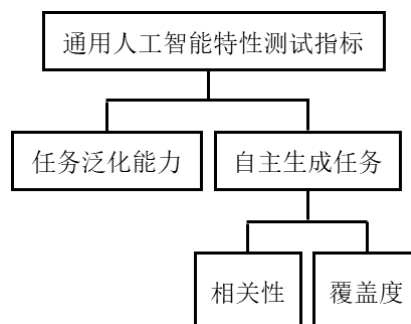


图1 家庭环境通用人工智能视觉理解能力测试指标

4.2 任务泛化能力

“任务泛化能力”指标用于评估AGI系统在家庭环境长时间运行中，对多任务类别的家庭环境视觉任务的泛化理解能力。衡量AGI系统在多任务场景下正确理解任务的精度，包括对已知任务和未知任务的泛化能力。计算方法见公式（1）：

$$T = \mu - \gamma \cdot \sigma \dots \dots (1)$$

$$\text{其中, } \mu = \frac{1}{M} \sum_{i=1}^M f_i, \sigma = \sqrt{\frac{1}{M} \sum_{i=1}^M (f_i - \mu)^2}$$

$$f_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \times 100\%, i \in \{1, 2, \dots, M\}$$

式中：

T——任务泛化能力

f——各任务类别的准确率

M——测试任务类别总数量

σ ——各任务类别准确率的标准差
 μ ——各任务类别准确率的均值
 γ ——任务表现波动性惩罚强度
 TP——正确完成的家庭任务数
 TN——正确忽略的家庭任务数
 FP——错误完成的家庭任务数
 FN——错误忽略的家庭任务数

4.3 自主生成任务

“自主生成任务”指标用于衡量AGI在无明确指令时，基于环境视觉输入自主发现、定义并执行合理任务的能力。对AGI自主生成任务的测试包括以下两种性质。

a) **相关性 (Relevance)**：测试生成任务与家庭场景实际需求的关联程度，避免无关或无效任务。计算方法见公式 (2)：

$$R = \frac{1}{M} \sum_{i=1}^N r(T_i, H) \dots\dots (2)$$

其中, $r(T_i, H) = \sum_{h_n \in H} \omega_j \cdot \cos(E(T_i), E(h_n))$

式中：

R —— 总任务集相关性

r —— 任务间相关性

N —— 测试周期内自主生成的任务总数

T_i —— 第i个自主生成的任务 (i取值范围1到N)

H —— 家庭场景的真实任务需求集合 (如清洁、整理、安全等典型家庭任务)

h_n —— H集合中的第j个真实家庭任务 ($h_n \in H$)

w_n —— 任务 h_j 的权重系数，反映该任务在家庭场景中的相对重要性 (所有 w_i 之和为1)

$E(\cdot)$ —— 任务文本的嵌入向量表示函数 (通常使用BERT等预训练语言模型获得)

$\cos(\cdot, \cdot)$ —— 余弦相似度函数，计算两个向量之间的相似度 (值域[-1, 1]，实际使用时取绝对值并归一化到[0, 1]范围)

b) **覆盖度**：衡量系统能够理解的家庭任务类别的多样性，反映泛化能力。计算方法见公式 (3)：

$$TC = \frac{|\Omega_{completed}|}{|\Omega_{total}|} \dots\dots (3)$$

式中：

TC——任务覆盖度

$\Omega_{completed}$ ——成功完成的任务类型集合

Ω_{total} ——测试中所有任务类型的集合

$|\cdot|$ ——集合的基数 (元素数量)

c) 自主生成任务：自主生成任务测试指标总分计算见公式（4）：

$$A = \alpha \cdot R + (1 - \alpha) \cdot TC \dots\dots (4)$$

A——自主生成任务指标得分
R——任务相关性
TC——任务覆盖度
α——任务相关性权重

5 测试任务维度

5.1 概述

家庭环境通用人工智能视觉理解测试任务维度主要分为静态视觉理解和动态视觉理解，见图 2。静态视觉理解包括客体属性及关系理解、视觉语言联合解译、状态理解、视觉常识理解 4 个二级维度。动态视觉理解包括事件理解、行为意图理解、情绪理解 3 个二级维度。视觉理解能力测试维度和典型任务见表 1。

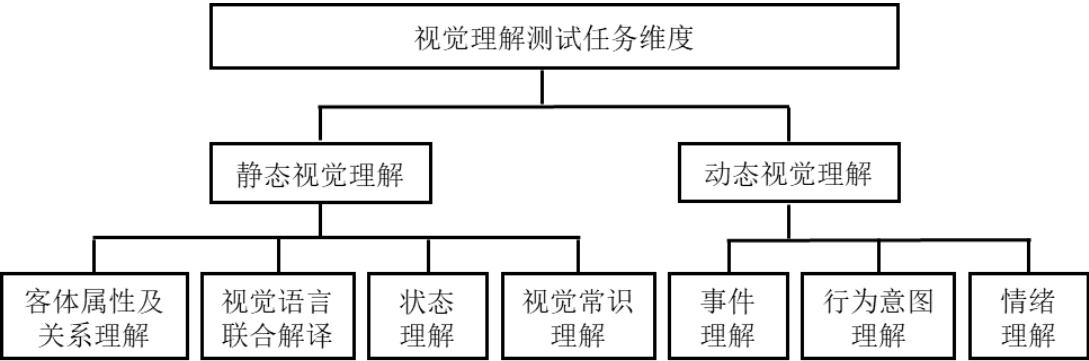


图2 家庭环境通用人工智能视觉理解能力测试指标

表1 理解能力评测维度和典型任务

一级维度	二级维度	描述	典型任务
静态视觉理解	客体属性及关系理解	系统能够识别出图像的核心内容，从图像中提取关键信息，确保识别结果具有较高的精确度和鲁棒性；能理解不同物体之间的空间关系。	对比被测对象输出图像识别对象与真实对象标签，如一致，则记为正确；图像深度估计、空间结构理解、三维形状理解；分析场景中物体间的空间、功能等关系（如上下关系、支撑关系、遮挡关系、铰链关系等）。
	视觉语言联合解译	模型能从提取图像中的文字，并联系图像与文本内容进行深入理解和分析。	提取图像文字、回答关于图像的文本问题、文档理解。
	状态理解	理解自身状态，以及环境的物理状态、属性或变化。	分析自身属性状态，判断物体材质、老化、损坏状态等。
	视觉常识理解	实现对常见的基本物理常识、社会常识的理解。物理	理解物体持久性（物品遮挡后依然存在）；理解重力与稳定性（如，积木堆

		常识包括基本的物理规律，社会常识包括社会规范、社交规则等。	叠的合理性)；理解材质交互规律(如，水杯打翻后液体流动)；理解因果关系。理解社交的基本规范；理解基础的文化符号(如，白色在西方是婚纱，在东方可能象征丧葬)。
动态视觉理解	事件理解	识别或预测图像或视频中的事件或动态变化。	基于视频理解事件，时序动作定位。
	行为意图理解	能识别图像或视频中人体的动作。推断人或智能体的目标、动机或交互意图。	识别视频中人体动作，如行走、跳跃、弯腰等；理解复杂人体行为，如打架、跌倒等；预测人类意图，识别人机交互意图。
	情绪理解	识别人脸表情或情绪状态。	表情识别、情绪分析和微表情理解。

5.2 客体属性及关系理解

测试客体属性及关系理解能力，指测试 AGI 识别图像中物体的属性(如颜色、形状、材质)及物体间的空间、逻辑关系的能力，包含但不限于以下能力：

- a) 属性识别：检测物体的物理或抽象特征，包含图像深度估计、空间结构理解、三维形状理解等；
- b) 关系推理：判断物体间的相对位置、归属关系或比较关系。

5.3 视觉语言联合解译理解

测试视觉语言联合解译能力，指测试 AGI 结合视觉与文本信息进行跨模态推理的能力，包含但不限于以下能力：

- a) 图文匹配：判断文本描述是否与图像内容一致；
- b) 视觉问答：回答基于图像内容的提问；
- c) 指代消解：关联文本中的代词或抽象表述到图像中的具体区域。

5.4 状态理解

测试 AGI 的状态理解能力，指测试 AGI 识别物体或场景的瞬时状态或条件的能力，包含但不限于以下能力：

- a) 物理状态判断：分析并准确判断物体的状态，如可操作性、完整性、开关状态、物理状态；
- b) 场景状态推断：判断环境的物理、社会属性，如整洁度、空间大小；
- c) 自身状态理解：判断智能体自身的属性、特征、状态。

5.5 视觉常识理解

测试 AGI 的视觉常识理解能力，指测试 AGI 基于常识对视觉内容进行合理推断的能力，包含但不限于以下能力：

- a) 隐含信息推理：推测图像中未直接呈现的信息；
- b) 反事实推理：预测违反常识的视觉情境。

5.6 事件理解

测试 AGI 的事件理解能力，指测试 AGI 识别并解析动态场景中的事件及其结构的能力，包含但不限于以下能力：

- a) 事件检测：识别定位视频中发生的特定事件；
- b) 时序关系建模：判断多个事件的时间顺序；
- c) 因果推理：分析事件间的因果关系。

5.7 行为意图理解

测试 AGI 的行为意图理解能力，指测试家庭环境 AGI 推断动态场景中主体的目标或动机的能力，包含但不限于以下能力：

- a) 动作意图识别：判断行为的潜在目的；
- b) 群体行为预测：预测多个主体间彼此协调或冲突的意图。

5.8 情绪理解

测试 AGI 的情绪理解能力，指测试 AGI 识别动态场景中主体的情绪状态的能力，包含但不限于以下能力：

- a) 面部情绪识别：通过表情判断情绪；
- b) 肢体语言分析：通过姿势或动作推断情绪；
- c) 情境情绪推理：结合上下文推测情绪。

6 测试流程

6.1 概述

AGI 视觉理解能力的测试包括测试准备、测试执行、测试结果三大步骤，如图3所示：

- a) 测试准备：将根据测试需要，筛选任务维度和任务池中的部分或全部任务作为测试任务集合，对每个任务用到的数据集进行质量审查；
- b) 测试执行：导入待测模型，将模型在测试任务集合上进行测试，测试逐维度、逐任务开展，直到全部选定任务均测试完成；
- c) 测试结果：根据待测模型在参加所有测试的测试数据，汇总报告测试结果。

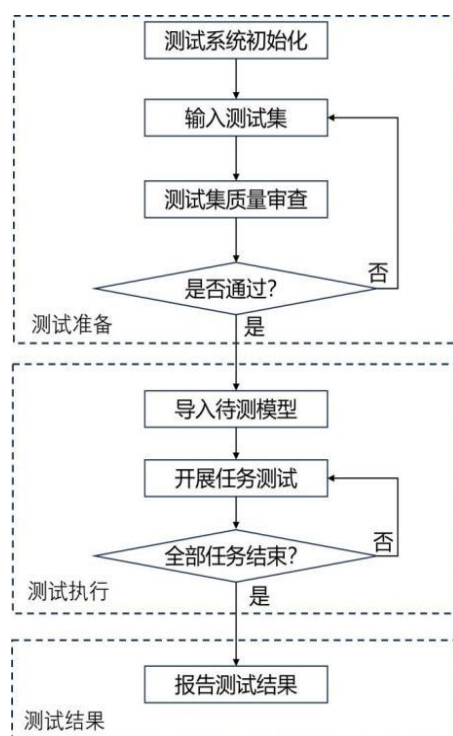


图3 AGI视觉理解能力测试流程

6.2 测试准备

测试数据集应满足以下要求：

- 合规性和隐私保护：数据收集过程遵循适用的法规和隐私保护标准，保护用户隐私；
- 评测指标完备：为每个评测指标构建满足相应数量的数据集；
- 时效性：数据集结合开源数据集和自制数据集，定期更新维护；
- 可用性：数据集格式和接口符合广泛的标准，以便于获取和使用；
- 多样性和代表性：涵盖不同的背景、场景、领域等，以确保数据能覆盖不同的使用情况；
- 数据标注流程符合GB/T 42755—2023中第6章和第7章的要求。

6.3 测试执行

测试执行过程包括以下三个步骤：

- 导入待测模型，验证模型的输入和输出数据工作正常；
- 逐个开展任务测试，由测试系统根据任务要求来记录待测模型表现；
- 汇总测试数据，进入测试结果分析及报告流程。

6.4 测试结果

6.4.1 “任务泛化能力”分数计算

该指标的计算包括以下步骤：

- 针对表1中的部分或全部二级能力维度，选择M个测试任务类别；
- 依次针对M个任务类别开展测试；
- 首先在选定的任务场景下，由测试系统随机生成针对选定任务类别的多个家庭任务，包含有效家庭任务和无关家庭任务；

- d) 给定明确任务指令，并设定任务执行时间，在时间用尽后汇总任务完成情况；
- e) 统计正确完成的家庭任务数（TP）、正确忽略的家庭任务数（TN）、错误完成的家庭任务数（FP）、错误忽略的家庭任务数（FN），根据公式1计算该类别得分；
- f) 综合全部M个类别得分，根据公式1计算“任务泛化能力”最终得分。

6.4.2 “自主生成任务”分数计算

该指标的计算包括以下步骤：

- a) 针对表1中的部分或全部二级能力维度，选择M个测试任务类别；
- b) 依次针对M个任务类别开展测试；
- c) 首先在选定的任务场景下，由测试系统随机生成针对选定任务类别的多个预期应完成的家庭任务；
- d) 进而为测量自主性，不设定明确任务指令，允许待测模型自主决定执行或不执行任意任务，但设定任务执行时间，在时间用尽后汇总任务完成情况；
- e) 统计所有已完成任务，根据公式2计算相关性，根据公式3计算覆盖度，最后根据公式4计算该类别“自主生成任务”得分；
- f) 综合全部M个类别得分，计算所有类别的平均分作为“自主生成任务”最终得分。

附 录 A
(资料性)
家庭环境通用人工智能视觉理解能力测试结果示例

A.1 测试任务

针对表1中的3个二级能力维度（“客体属性及关系理解”、“视觉语言联合解译”、“视觉常识理解”），选择8个家庭环境测试任务，分别是整理行李箱、整理桌面、拼图、搭建积木、物品分类计数、打扫房间、理解按钮功能、挑选礼物。

具体来说，整理行李箱、整理桌面、拼图、搭建积木这4个任务对应“客体属性及关系理解”维度；物品分类计数、打扫房间这2个任务对应“视觉语言联合解译”维度；理解按钮功能、挑选礼物对应“视觉常识理解”维度。基于UnrealEngine引擎开发了虚拟测试场景，支撑上述8个家庭任务的测试。测试场景如下：



图A.1 测试任务场景示意图

针对每一项测试任务，检查所采用的虚拟测试环境是否能够正常启动，导入智能体，完成正常的数据交换。并且按照6.2中要求，对所采用的测试环境进行检查。如通过检查，则进入下一步测试。

A.2 测试结果

实验中接入2款多模态大模型 GPT-4o-2024-08-06（简称为 GPT-4o）、Claude-3.5-sonnet-20240620（简称为 Claude3.5）。模型在各项测试任务中的原始得分如下表所示（百分制）。可以根据公式1，进一步计算均值和标准差，从而得到“任务泛化能力”指标的分数。当 $\gamma=1$ ，GPT-4o的“任务泛化能力”得分为0.172，而Claude3.5的得分为0.108。

表A.1 模型测试结果

模型	整理行李 箱	理解按钮 功能	整理桌 面	挑选礼 物	物品分类 计数	搭建积 木	打扫房 间	拼图
GPT-4o	55%	40%	30%	29%	33%	13%	24%	15%
Claude3.5	5%	33%	13%	34%	27%	20%	15%	16%

参 考 文 献

- [1] GB/T 41867-2022 信息技术 人工智能 术语
 - [2] YD/T 4316-2023 面向智慧城市应用的人工智能服务能力开放技术要求
 - [3] YD/T 3944-2021 人工智能芯片基准测试评估方法
 - [4] GB/T35273—2020 信息安全技术 个人信息安全规范
 - [5] GB/T40660—2021 信息安全技术 生物特征识别信息保护基本要求
 - [6] GB/T45225—2025 人工智能 深度学习算法评估
 - [7] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程 英文标准名称
 - [8] T/CIE 298-2025 人形机器人智能化分级
-

附件 2

意见反馈表

提出单位（公章）： _____

联 系 人： _____ 联系电话： _____

传 真： _____ E - mail: _____

序号	章条编号	修改意见内容（包括理由或依据）

注：如所提意见篇幅不够，可增加附页。