ICSxx.xxx.xx
CCS X xx

团 体 标 准

JH/CAA 007-2025

无人系统关键部组件智能测评指南

Guidelines for intelligence evaluation of key components in unmanned systems

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上

2026-XX-XX 发布 2026-XX-XX 实施

中国自动化学会 发 布

目　　次

前　言

本文件按照GB/T1.1—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。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国自动化学会提出并归口。

本文件起草单位：启元实验室、江淮前沿技术协同创新中心、哈尔滨工业大学、……。

本文件主要起草人：

无人系统关键部组件智能测评指南

1 范围

本文件给出了无人系统关键部组件智能测评的总体设计、需求分析、测评设计、测评执行和能力评估方案。

本文件适用于对无人系统（如无人机、无人车、机器人等）中具备智能性的关键部组件，以及对无人系统智能性具有直接贡献的关键部组件智能特征的测试和评估。

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。

ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

3 术语和定义

ISO/IEC 22989:2022界定的以及下列术语和定义适用于本文件。

3.1

无人系统 unmanned system

能够在无人操作的情况下，依靠自身控制与决策能力执行指定任务的机电系统。

3.2

关键部组件 key component

自身具备一定智能能力或对整个系统智能具有贡献的部件或组件。

3.3

智能测评 intelligence evaluation

对关键部组件的智能特征相关指标进行测量、分析并综合评估的过程。

3.4

功能原子 function element

关键部组件能够独立实现的最小智能能力单元。

3.5

感知能力 perceptual ability

关键部组件对环境信息进行获取、处理和理解的能力。

3.6

决策能力 decision-making ability

关键部组件在面对不同情境、条件或选择时做出有效决策的能力。

3.7

执行能力 execution ability

关键部组件将决策、规划或操作转化为行动的能力。

3.8

自主性 autonomy

能够在没有外部干预、控制或监督的情况下修改其预期使用领域或目标的系统的特性。

[来源：ISO/IEC 22989:2022，3.1.5，有修改]

3.9

适应性 adaptability

关键部组件在面对环境、任务或输入数据变化时，维持或快速恢复其性能水平的特性。

3.10

学习性 learnability

关键部组件在其全生命周期的运行阶段，持续进行增量训练提升其任务执行效果与应对复杂场景能力的特性。

[来源：ISO/IEC 22989:2022，3.1.9，有修改]

3.11

可信性 trustworthiness

关键部组件在开放、动态和不确定的环境中，持续表现出其行为符合人类预期、符合技术规范、符合伦理法律准则，从而能够获得人类用户、监管机构及社会公众信任的综合特性。

3.12

可解释性 explainability

关键部组件以人类能够理解的方式表达影响人工智能系统结果的重要因素的特性。

[来源：ISO/IEC 22989:2022，3.5.7，有修改]

3.13

单元测试 unit test

对系统中的关键部组件进行隔离和验证的测试过程。

3.14

集成测试 integration test

将已经通过单元测试的各个部组件按照设计要求组装成一个较大的系统或子系统，然后对这个组装后的整体进行测试的过程。

3.15

虚实结合测试 virtual-real combination test

在测试过程中，将虚拟环境（如仿真模型、虚拟场景等）与真实环境（如物理设备、实际场景等）相结合，通过实时交互和数据比对，对关键部组件进行功能、性能、可靠性等方面的验证和评估。

4 测评总体设计

4.1 测评原则

4.1.1 核心能力聚焦原则

关键部组件智能测评无需覆盖感知、决策、执行全部能力，宜基于其设计目标与应用场景，聚焦于其主要贡献的智能特征进行测评。

示例：视觉感知算法的测评应重点针对其感知能力，而非决策或控制能力。

4.1.2 保守评价原则

在指标评估存在不确定性或多结果并存时，宜采纳保守的评价结论，避免对部组件性能进行过高或乐观的估计。

4.1.3 应用场景驱动原则

制定具体部组件的测评方案时，宜首先明确其应用场景，根据场景特点选取和深化相应的测试维度和指标，从而确保测评结果能够真实、有效地反映该部组件在其特定应用场景下的智能水平。

4.1.4 标准优先原则

测评过程中，针对具体指标的测试方法，宜优先检索并采用强制性国家标准规定的测试方法；如无强制性标准，则宜依次采纳推荐性国家标准、行业标准中的方法；在上述标准均缺失时，方可使用经权威机构认可的内部方法或公认的通用技术方法。

4.2 测评流程

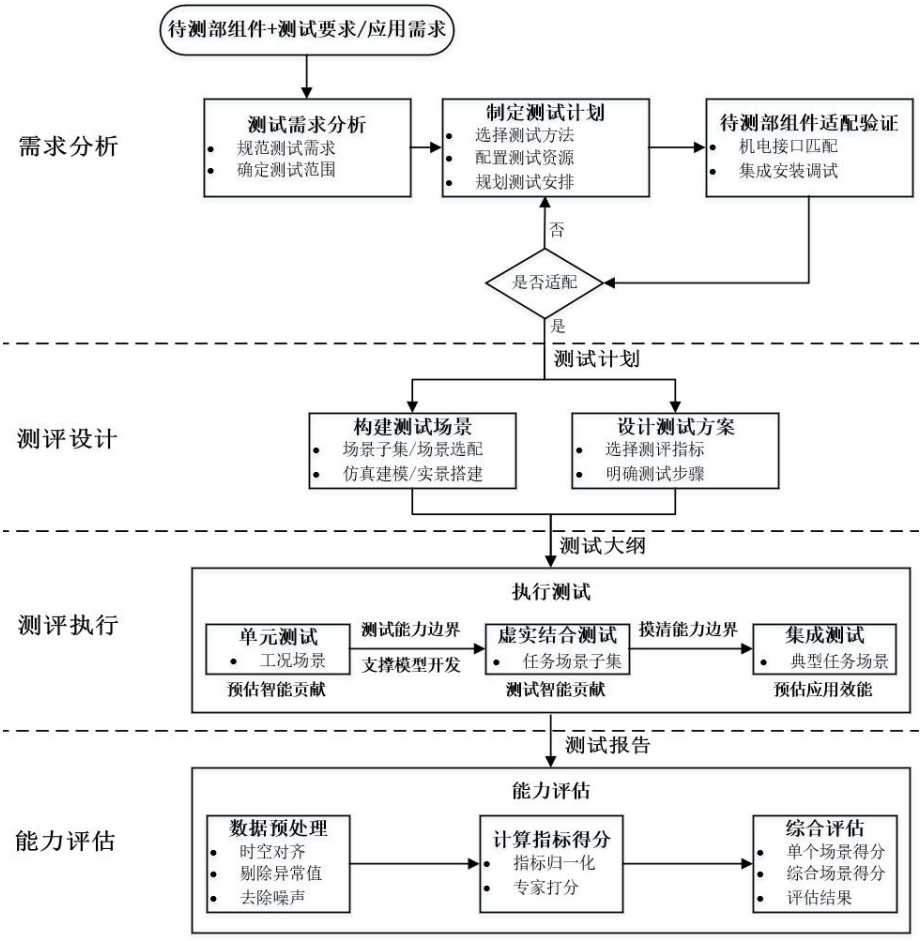

图 1 部组件智能测评流程

为科学量化评价无人系统关键部组件在典型任务场景下的真实性能与智能贡献，制定如图1的测评流程：

a)需求分析阶段：首先，根据待测部组件及其测试要求或应用需求进行测试需求分析，确定测试范围。接着，制定测试计划，选择测试方法并配置测试资源。然后，对组件进行适配验证，确保其与测试环境兼容。此阶段的输入包括部组件介绍、运行设计条件等，输出为测试需求描述和计划。

b)测评设计阶段：基于需求分析的结果，构建测试场景和设计测试方案。具体包括选择测评指标、明确测试步骤以及场景建模或实景搭建。输入为测试需求描述和计划，输出为测评指标、场景集等。

c)测评执行阶段：按照测试大纲，依次进行单元测试、虚实结合测试和集成测试，逐步验证各部分功能和整体性能。输入为测试大纲，输出为测试数据、过程记录和报告。

d)能力评估阶段：对测试结果进行综合评估，包括数据预处理、计算指标得分和综合评估。最终形成测试报告，展示各项指标的测试结果和综合场景得分。

5 需求分析

需求分析是智能测评的起点，旨在明确测评目标与路径，为后续活动提供依据。本阶段包括测试需求分析、制定测试计划与待测部组件适配验证。

a) 测试需求分析：基于部组件设计文档与应用场景，规范其核心智能特征与性能要求，形成明确、可测试的需求条目。据此确定测评范围，包括需覆盖的智能能力维度、待测的功能原子、测试类型（单元、虚实结合、集成）及场景层级（代表性、预期、非预期）。

b) 制定测试计划：依据测试需求，制定全局性测试计划。选择测试方法，优先采用国家标准或行业标准方法。配置测试资源，明确人员职责、设备清单（仿真平台、测试台架等）及测试场地。规划测试安排，制定进度表并进行风险评估，确保测试有序、可控。

c) 待测部组件适配验证：在测试开始前，确保部组件与测试环境兼容。完成机电接口匹配检查与物理连接。进行集成安装调试，验证通信链路、驱动基础功能并完成传感器标定与时空同步，确认部组件已就绪。

6 测评设计

6.1 测试场景的设计

6.1.1 测试场景分级

测评系统宜通过构建多层次的场景集，既能验证部组件的常规性能，亦可揭示其在非预期挑战下的脆弱性，为优化与安全部署提供关键依据。

测试场景宜分为三个层级：

a) 基础场景：选取具备高曝光度与应用价值的典型任务，用以验证系统核心功能并建立性能基准。

b) 挑战场景：通过参数变异与要素组合，衍生出已知的挑战性场景，检验系统在已知扰动下的适应性与学习性等。

c) 边界场景：主动构建包含未知扰动、极端条件或模拟故障的场景，探测系统的极限容错能力与潜在安全隐患。

6.1.2 测试场景构建

对于关键部组件的测试场景构建应综合考虑以下四类要素，其关系如图2所示：

a) 无人系统配置：包括载体平台、性能指标、协作系统等。

b) 外部环境：包括泛自然地理层、设置层、动态层、数据操作层等。

c) 任务设定：包括任务初始状态、任务过程、任务目标等。

d) 临时事件：包括无人系统相关事件（如突发故障）、外部环境相关事件（如通信干扰）、任务设定相关事件（如指令变更）等。

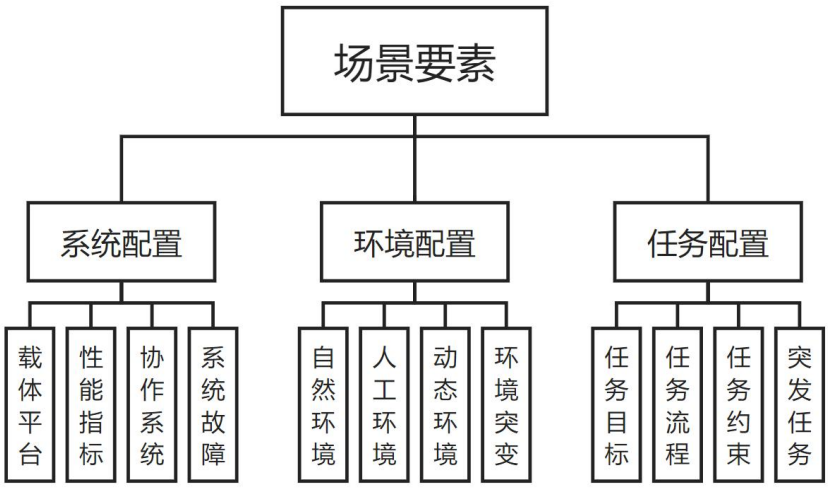

图 2　场景要素设计框架

6.1.3　测试场景要求

a) 覆盖充分性:场景集应能系统性地覆盖第6.1.1条定义的代表性场景、预期场景和非预期场景三个层级。场景复杂度应呈梯度分布，以全面测评部组件从基准性能到边界性能的表现。

b) 物理合理性与真实性:虚拟场景及仿真模型应具备足够的保真度，其动力学、传感器、环境效应等模型应能反映真实物理世界的规律。

6.2　测试指标的选取

智能的表现应涵盖以下八个维度：广泛的能力边界、高度的准确性、卓越的效率、可靠的稳定性、良好的自主性、强大的适应性、持续的学习能力以及安全可信的运行保障。

由于部分部组件功能交叉且难以直接比较，宜将其能够独立实现的最小智能能力单元抽象为“功能原子”。功能原子是测评设计的基本对象，用以清晰表征各部组件对系统智能的具体贡献。

基于无人系统的智能能力（感知、控制决策、执行），分析部组件在其中所承担的角色，将其核心能力分解为若干个可独立测试的功能原子。部组件功能原子与系统智能能力的映射关系示例如图3所示。

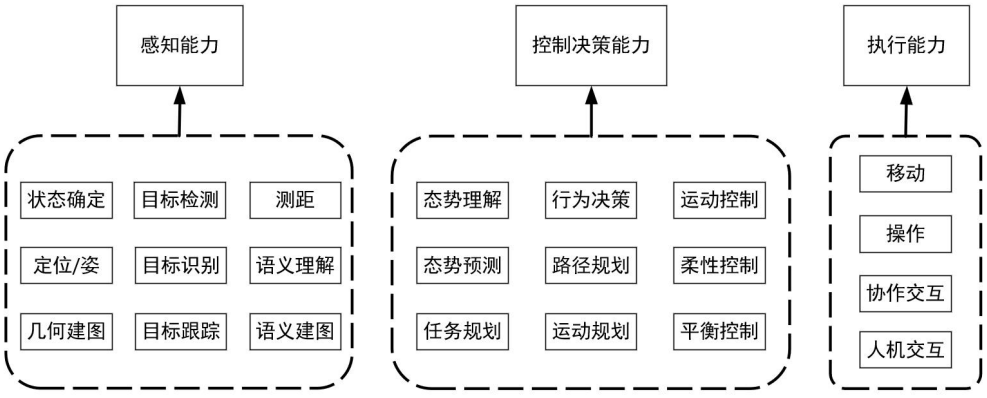

图 3　关键部组件功能原子实例

针对各部组件的每一项功能原子及其对应的智能特征（感知-决策-执行），具体测评时宜根据

4.1.1 原则，参考表 1 所示的智测指标共性设计框架，仅选取与被测部组件核心智能特征相关的指标。

表 1　智能特征评价指标

测评维度		具体指标举例	单元测试	虚实结合测试	集成测试	关键说明与理由
能力与性能层	能力边界	最大探测距离、最大处理目标数、最大负载等	★	✓	△	单元测试在台架上可精准找到性能拐点。虚实结合可安全地探索边界，但模型保真度影响结果。集成测试风险高，常用于验证。
	准确性	感知准确率、定位精度、控制误差等	★	✓	★	单元测试用标准数据集/高精度设备测算法/部件本身。虚实结合受"现实差距"影响，适于相对比较。集成测试获取系统在真实环境下的最终精度。
	效率	算法耗时、吞吐率、CPU/GPU 占用等	★	✓	★	单元测试适用于测量隔离环境下的纯计算效率。虚实结合测试适于算法 A/B 测试，但绝对时间可能失真。集成测试适用于测量任务级的总时间和能耗。
	稳定性	短期输出波动、长期漂移、零偏稳定性等	★	★	✓	单元测试适用于测量物理器件的固有稳定性（如传感器零偏）。虚实结合测试适用于测试算法/软件在重复和扰动下的逻辑稳定性。集成测试能验证部组件在真实扰动下的长期稳定性，但缺点是难以区分性能波动是来自被测部组件，还是系统其他部分，仅用于补充分析。
智能特征	自主性	人工干预频率、自我诊断、能量管理等	△	✓	★	自主性是系统级属性。单元测试几乎不适用。虚实结合可在复杂场景中初步评估。集成测试在真实任务中给出最终评价。
	适应性	对光照/天气/干扰的性能保持率、恢复速度等	✓	★	✓	适应性需要在变化环境中测试。单元测试和虚实结合是可重复、系统化测试适应性的主力战场，尤其擅长故障注入。集成测试可重复性差、成本高，难以系统性覆盖所有预期扰动。
	学习性	性能提升率、学习效率等	△	★	✓	学习性需要在大量、多样的数据/场景中评估。虚实结合能提供近乎无限、可控的数据源，是测试学习性的理想平台。集成测试难以提供学习性测试所需的大规模、多样化、可控的数据，通常只用于验证学习后的最终效果。
	可信性	可靠性、安全性、伦理性、可解释性等	△	★	★	虚实结合测试适用于可靠性/安全性的故障注入和极端场景测试。伦理性/可解释性需结合仿真场景与专家评估。最终的系统级可信性必须在集成测试中确认。

测评维度	具体指标举例	单元测试	虚实结合测试	集成测试	关键说明与理由
注： ★（主要）：该测试方法是获取此项指标最直接、最准确、最权威的手段。 ✓（辅助）：该测试方法可以用于评估、比较或验证此项指标，但存在局限，或通常作为辅助和前期验证手段。 △（有限）：该测试方法在此项指标上能力非常有限，通常无法提供可靠数据。					

7 测评执行

7.1 能力边界测试

能力边界测试宜在可控条件下(如测试台架)，逐步增大输入或任务难度，直到部组件性能失效或急剧下降至不可接受水平。能力边界的评估指标宜包括但不限于以下。

性能拐点：当性能低于临界阈值$P_{critical}$时对应的输入值，视为达到能力边界，参考图 4。

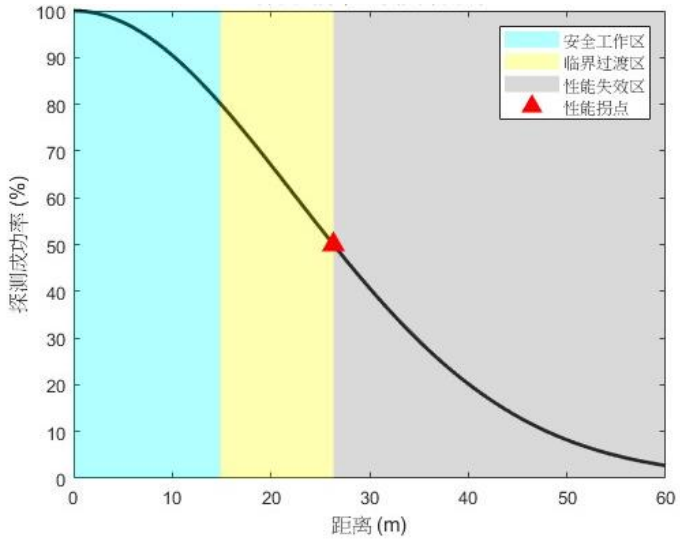

图 4 能力边界分析实例（探测成功率-距离曲线）

7.2 准确性测试

准确性测试宜在高精度测量设备提供的“真值”参考下，进行大量重复测量，并进行统计分析。准确性的评估指标宜包括但不限于以下。

a) 平均绝对误差，计算公式为：

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\| Y_i-Y_{true,i}\| \quad (1)$$

式中，

N —— 总测量次数；

Y_i —— 第 i 次测量的测量值；

$Y_{true,i}$ —— 第 i 次测量的真值。

b)最大/最小误差：误差绝对值的最大值与最小值。

对于分类检测等功能原子，其准确性的评估指标宜包括以下。

a)准确率，计算公式为：

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

b)精确率(查准率)，计算公式为：

$$P=\frac{TP}{TP+FP} \quad (3)$$

c)查全率(召回率、反馈率)，计算公式为：

$$R=\frac{TP}{TP+FN} \quad (4)$$

d)F1-Score，计算公式为：

$$F_1=2\times\frac{P\times R}{P+R} \quad (5)$$

式中，

TP —— 真正样本(True Positive)；

TN —— 真负样本(True Negative)；

FP —— 假正样本(False Positive)；

FN —— 假负样本(False Negative)。

7.3 效率测试

效率测试宜在标准任务和负载下，监测其资源消耗情况。效率测试的评估指标宜包括但不限于以下。

a)任务完成时间，计算公式为：

$$T_{total}=T_{end}-T_{start} \quad (6)$$

b)平均任务完成时间，计算公式为：

$$\overline{T}=\frac{1}{N}\sum_{i=1}^{N}T_{total,i} \quad (7)$$

c)吞吐率(though put rate)，计算公式为：

$$TP=\frac{N}{T_{total}} \quad (8)$$

d)响应时间，计算公式为：

$$T_{response}=T_{first\_output}-T_{input\_arrival} \quad (9)$$

式中，

T_{start} —— 任务开始的时刻；

T_{end} —— 任务成功完成的时刻；

N —— 任务次数；

$T_{total,i}$ —— 第 i 次任务的任务完成时间；

N_{tasks} —— 在T_{total}时间内完成的任务总数;

$T_{response}$ —— 系统从接收输入到产生输出所经历的时间;

T_{first_output} —— 系统产生第一个输出的时刻;

$T_{input\_arrival}$ —— 系统从接收输入到产生输出所经历的时间。

7.4 稳定性测试

7.4.1 短期稳定性测试

短期稳定性测试宜在相同条件下，于较短时间内（分钟/小时尺度）多次运行，采集记录数据（如精度指标）。短期稳定性测试的评估指标宜包括但不限于以下。

a)标准差(Standard Deviation,SD)计算公式为：

$$SD=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2} \quad (10)$$

式中，

Y_i —— 第 i 次测量值；

N —— 总测量次数；

$\bar{Y}$ —— 测量值的均值。

7.4.2 长期稳定性测试

长期稳定性测试宜在可控环境下，进行长时间持续监测，记录输出数据。长期稳定性测试的评估指标宜包括但不限于以下。

a)线性漂移率，计算公式为：

$$D_{linear}=\frac{Y_{end}-Y_{start}}{T_{total}} \quad (11)$$

式中，

Y_{end} —— 长时间测试结束时的主要性能指标输出值；

Y_{start} —— 长时间测试开始时的主要性能指标输出值；

T_{total} —— 总测试时长。

b)趋势漂移率：对输出-时间数据进行线性回归，斜率即为趋势漂移率

d)输出范围/峰峰值，计算公式为：

$$Y_{pp}=Y_{max}-Y_{min} \quad (12)$$

式中，

Y_{max} —— 在测试期间记录到的性能指标的最大值；

Y_{min} —— 在测试期间记录到的性能指标的最小值。

e)变异系数，计算公式为：

$$CV=\frac{SD}{\bar{Y}}\times 100\% \quad (13)$$

式中，

SD —— 测量值的标准差；

$\bar{Y}$ —— 测量值的均值。

7.5 自主性测试

自主性测试宜在代表性场景中进行任务级测试，记录关键数据。自主性测试的评估指标宜包括但不限于以下。

a)人工干预频率@任务完成度：单位时间或单位任务（如每百公里驾驶、每百次操作）中需要人类接管的次数。例：4@100%，表示任务完成度 100%情况下人工干预 4 次

b)人工干预时长/比例@任务完成度：人工干预时长占总任务时长的比例。例：20%@80%，表示任务执行时间 20%下由人工干预，最终任务完成度为 80%。

c)自我诊断能力，计算方法为：

$$\text{自我诊断能力}=\begin{cases}1 & \text{如果系统具备自我诊断能力}\\ 0 & \text{如果系统不具备自我诊断能力}\end{cases} \quad (14)$$

d)能量管理能力，计算方法为：

$$能量管理能力=\begin{cases}1 & 如果系统具备能耗管理与节能功能\\0 & 如果系统不具备能耗管理与节能功能\end{cases} \quad (15)$$

7.6 适应性测试

7.6.1 静态环境适应性测试

静态环境适应性宜在不同环境参数（光照、天气、地形、干扰物分布、交通密度等）的静态代表性场景下测试，与基线（代表性场景）性能对比。静态环境适应性测试的评估指标宜包括但不限于以下。

a)性能保持率(Performance Retention Rate, PRR)，计算方法为：

$$PRR=\frac{AP-BP}{BP} \quad (16)$$

式中，

AP —— 预期环境下的实测性能（准确率、成功率等）；

BP —— 基线环境（代表性场景）下的性能。

b)性能下降绝对值，计算方法为：

$$\Delta P = AP - BP \quad (17)$$

c)恢复速度，计算方法为：

$$T_{recov} = T_{stable} - T_{perturb} \quad (18)$$

式中，

T_{stable} —— 性能指标首次进入并保持在“稳定阈值”范围内的时刻；

$T_{perturb}$ —— 环境发生突变或干扰开始的时刻。

7.6.2 动态环境适应性测试

动态环境适应性宜在环境参数连续或阶跃变化的场景中测试，记录性能实时曲线。（如：天气从晴→雨→雾；交通流从稀疏→拥堵）。动态环境适应性测试的评估指标宜包括但不限于以下。

a)滞后性，环境发生阶跃变化(T_{change})到系统性能开始显著下降(T_{drop})的时间差，计算方法为：

$$T_{lag} = T_{drop} - T_{change} \quad (19)$$

式中：

T_{drop} ——性能指标首次持续低于基线性能$P_{baseline}$的 95%的时刻；

T_{change} ——环境发生阶跃变化的时刻。

b)恢复时间，性能从开始下降(T_{drop})到恢复至稳定状态(T_{stable})所需的时间，计算方法为：

$$T_{recov} = T_{stable} - T_{drop} \quad (20)$$

式中：

T_{stable} ——性能指标首次持续低于基线性能$P_{baseline}$的 95%的时刻；

T_{drop} —— 性能从开始下降时刻。

c)性能损失，性能在恢复过程中相对于基线性能的最大下降值，计算方法为：

$$P_{loss} = P_{baseline} - P_{min} \quad (21)$$

式中：

$P_{baseline}$ —— 基线性能；

P_{min} —— $[T_{drop}, T_{stable}]$时间区间内的性能最小值。

d)振荡幅度，在性能恢复过程($[T_{drop}, T_{stable}]$)中，性能值的波动程度。用该时间段内性能数据的标准差来衡量，计算方法为：

$$s = StdDev(\{P_t \mid t \in [T_{drop}, T_{stable}]\}) \quad (22)$$

7.6.3 压力适应性测试

压力适应性测试宜在系统运行时，人为注入传感器噪声、数据丢失、执行器偏差、通信延迟等故障。记录性能的瞬时下降幅度、故障消除后的自恢复时间和效果。压力适应性测试的评估指标宜包括但不限于以下。

a)性能衰减度：注入故障后的性能下降程度,该值越大，表明故障对系统性能的冲击越大，计算方法为：

$$PAD = \frac{(P_{baseline} - P_{min})}{P_{baseline}} \times 100\% \quad (23)$$

式中：

$P_{baseline}$ —— 基线性能；

P_{min} —— $[T_{drop}, T_{stable}]$时间区间内的性能最小值。

b)故障恢复时间：性能从故障消除(T_{remove})到恢复至稳定状态(T_{stable})所需的时间，计算方法为：

$$T_r = T_{stable} - T_{remove} \quad (24)$$

式中：

T_{stable} —— 性能指标首次进入并维持在$P_{baseline}$的 ±5% 范围内的时刻；

T_{remove} —— 故障消除时刻。

7.7 学习性测试

7.7.1 性能提升测试

性能提升测试宜在相同任务场景中进行固定次数的多次测试。性能提升测试的评估指标宜包括但不限于以下。

性能提升率(Performance Improvement Rate, PIR)，计算方法为：

$$PIR = \frac{P_{end} - P_{initial}}{P_{initial}} \quad (25)$$

式中：

$P_{initial}$ —— 学习前的性能；

P_{end} —— 学习后的性能。

7.7.2 学习效率测试

学习效率测试宜在相同任务场景中进行测试，不固定测试次数，记录性能曲线。学习效率测试的评估指标宜包括但不限于以下。

学习效率(Learning Efficiency, LE)，达到特定性能提升所需的学习次数，计算方法为：

$$LE = N_{iter} @ P_{target} \quad (26)$$

式中：

N_{iter} —— 达到目标性能所需的学习迭代次数；

P_{target} —— 指定的目标性能水平。

7.8 可信性测试

7.8.1 概述

可信性测试旨在评估关键部组件在实现其特定功能原子时，在安全、可靠、合规及可追溯等方面的表现。测试应聚焦于部组件接口层面的行为特征及其在系统故障链中的角色，重点关注在异常条件下的表现边界、失效模式可预测性以及支持系统级安全审计的能力。

7.8.2 可靠性测试

可靠性测试宜在典型工况下进行规定时长的连续运行，记录性能指标变化。评估部组件在长时间运行或重复任务下的性能稳定性和故障率。可靠性测试的评估指标宜包括但不限于以下。

a)平均故障间隔时间(MTBF)，系统两次相邻故障之间的平均工作时间，计算方法为：

$$MTBF = \frac{T_{total}}{N_{failure}} \quad (27)$$

式中：

T_{total} —— 总累计运行时间;

$N_{failure}$ —— 在总运行时间内发生的故障总次数。

b)可靠度(R(t))，系统在规定的条件和时间(t)内无故障运行的概率，计算方法为：

$$R(t)=P(T>t) \quad (28)$$

式中：

t —— 规定的时间;

T —— 产品的寿命。

7.8.3 安全性测试

安全性测试宜通过系统性故障注入、恶意攻击模拟（渗透测试）与风险分析，综合量化无人系统及关键部组件预防、抵御危险及从危害中恢复的能力。安全性测试的评估指标宜包括但不限于以下。

a)故障容忍时间，无人系统关键部组件在发生特定故障后，从故障发生到进入不可逆危险状态的最短时间，计算方法为：

$$FIT=T_{hazardous}-T_{fault} \quad (29)$$

式中：

$T_{hazardous}$ —— 进入危险状态的时间点;

T_{fault} —— 故障发生的时间点。

b) 对抗样本攻击成功率(ASR)，无人系统及其关键部组件成功识别的攻击尝试占总攻击尝试的比例。

$$ASR=\frac{TP}{TP+FP} \quad (30)$$

式中，

TP —— 真正样本(True Positive)；

FP —— 假正样本(False Positive)；

7.8.4 公平性测试

公平性测试宜构建包括不同属性特征的测试集，在相同的测试环境和条件下，使用部组件分别处理各属性子集的测试数据，记录部组件在每个子集上的核心性能指标并分析性能差异。公平性测试的评估指标宜包括但不限于以下。

a) 群体间性能极差：部组件在所有属性子集上，某一核心性能指标（如召回率）的最大值与最小值之差。计算公式为：

$$P_{range}=\max(P_i)-\min(P_i) \quad (31)$$

式中：

P_i —— 部组件在第 i 个属性子集上的性能指标值。

b) 群体间性能变异系数：部组件在所有属性子集上某一性能指标的标准差与其均值的比值，用于衡量差异的相对程度。计算公式参考式(13)。

7.8.5 可解释性测试

可解释性测试宜向部组件输入一组边缘案例（如高度遮挡目标、罕见场景模拟数据）或经过确认的对抗样本，检查部组件提供的解释性输出（如目标检测的置信度分数与类别概率分布；决策动作的候选选项列表及其效用评分或关键影响因素；规划路径的代价地图或风险热图）是否能合理解释其最终输出（无论对错），由领域专家根据解释性输出与最终输出的逻辑一致性进行判断。可解释性测试的评估指标宜包括但不限于以下。

解释有效性(V_{expl})：无人系统及其关键部组件提供的解释能准确反映其最终输出的比率

$$V_{expl}=\frac{N_{valid}}{N_{total}} \quad (32)$$

式中，

V_{expl} —— 解释有效性；

N_{valid} —— 专家判定部组件的解释性输出能够合理解释其最终输出的测试案例数；

N_{total} ——总测试案例数。

7.8.6 可追溯性测试

可追溯性测试宜定义一组部组件应记录的关键内部事件清单（例如：模块初始化完成/失败、自检告警、输入数据有效性检查异常、核心算法迭代收敛状态、输出置信度低于安全阈值、触发降级模式等）。在测试运行过程中，通过正常操作、注入故障或制造特定条件，确保清单中的每类事件均有发生。检查部组件的日志输出流或专用状态接口，验证是否每条发生的事件都被及时、准确地记录，且记录包含必要信息（如时间戳、事件 ID、严重等级、相关数据快照或上下文）。可追溯性测试的评估指标宜包括但不限于以下。

关键事件日志捕获率($R_{capture}$)：在测试中实际发生的、且属于预定义清单内的关键事件，被部组件成功记录并输出的数量占比。

$$R_{capture}=\frac{N_{logged}}{N_{occurred}} \quad (33)$$

式中，

N_{logged} —— 在实际发生的预定义关键事件中，被部组件成功记录并输出（可在日志中检索到完整记录）的事件数量；

$N_{occurred}$ —— 在测试过程中，实际发生且属于预定义关键事件清单内的事件总数。

7.9 测评实例

附录A提供了一个无人系统关键部组件测评的典型实例，以视觉感知部组件为例，展示了智能测评的全流程。

7.10 数据预处理

在计算评估指标前，应对测评执行阶段采集的原始数据进行预处理，以确保数据质量，消除非系统误差对评估结果的影响。预处理主要包括以下环节：

a）时空对齐：将来自不同传感器、不同时间戳或不同坐标系的数据统一到同一时间基准和空间坐标系下。例如，对于集成测试中的轨迹数据，需将惯性导航单元(IMU)、全球定位系统(GPS)和视觉里程计的输出进行时间同步和坐标变换，以确保数据关联的正确性。

b）剔除异常值：识别并移除因传感器瞬时故障、外部突发干扰或数据传输错误导致的明显不合理数据点（可采用统计方法如3σ准则或基于物理约束的门限法进行识别）。移除异常值后，宜采用插值法补充数据序列，或注明该时段数据无效。

c）去除噪声：采用合适的滤波算法对数据进行平滑处理，以抑制随机噪声，提取有效信号趋势。根据数据特性，可选用移动平均滤波、卡尔曼滤波或低通巴特沃斯滤波器等。滤波器的参数选择不应改变数据的本质特征，尤其应保留用于稳定性、适应性分析的高频动态信息。

8 能力评估

8.1 计算指标得分

基于预处理后的数据，参考第7章定义的各项测评指标，进行标准化处理，以支持跨维度比较。

a) 指标计算：参考第7章中定义公式的方法，逐一计算每个测评维度下的具体指标值（如准确性下的F1分数、效率下的响应时间等）。

b) 指标归一化：由于各指标量纲和物理意义不同，需将其统一映射到[0,1]区间或百分制，以消除量纲影响。归一化方法应根据指标特性选择：

对于高优类指标（正向指标，指标值越大越好，如准确率、MTBF）：归一化分值=(实测值-最差值)/(基准值-最差值。其中，基准值可为基准件性能或理论最优值，最差值可为性能容忍下限或理论最差值。

对于低优指标（负向指标指标值越小越好，如误差、耗时）：归一化分值 =（最差值-实测值)/(最差值-基准值)。

c) 专家打分：对于可信性等维度中难以直接量化的指标（如伦理性、可解释性），宜组建跨领域专家小组，参考第7章中提供的打分表（如表3、表4、表5）进行背靠背打分，最终取平均分或协商一致后的分数作为该指标得分。

8.2 综合评估

将各维度的归一化得分进行综合，形成对部组件智能水平的整体评价。

a) 单个场景得分：在单个测试场景下，部组件的综合表现可由其在所有测评维度上的得分的加权和表示。权重应根据部组件的功能定位和应用场景需求确定。例如，对于感知部组件，准确性和效率的权重可设置较高；对于决策部组件，自主性和可信性的权重可设置较高。若无特殊要求，可采用等权重计算。

b) 综合场景得分：部组件的最终评估应基于其在所有测试场景（代表性场景、预期场景、非预期场景）下的表现。可计算其在所有场景下得分的平均值，或采用“木桶原理”，重点关注其在非预期场景下的最低得分，以评估其性能下限和鲁棒性。

c) 基于基准件的偏序集比较与可视化：

偏序比较：将部组件在八个测评维度上归一化后的结果，与事先选定的基准件（代表行业平均水平或特定对标产品的部组件）在对应维度上进行逐一量化比较，形成“优于”、“持平”或“劣于”的客观偏序关系。

可视化呈现：将比较结果绘制于雷达图中，形成该部组件的“智能特征轮廓图”（如图5所示）。该图能直观展示其相对于基准件的能力强项与短板。

综合评估结论：基于雷达图进行综合评估，不仅能判断其与基准件的整体优劣，更能精准定位其技术成熟度与后续研发迭代的优先方向。

9 特殊要求

9.1 总体要求

无人系统关键部组件的测评执行应充分考虑载体平台的运动特性、环境约束与任务使命差异。不同类型无人系统在测试环境搭建、边界条件设定、测评侧重点等方面具有特殊要求。

9.2 无人机系统用部组件测评特殊要求

无人机系统对部组件的重量、功耗、空中感知与决策响应速度有严苛要求，测评时应予以侧重.

a） 测试场景特殊设计

感知能力：应增加对空中动态目标（如其他无人机、鸟类）的探测、跟踪与避让能力测试场景。视觉感知部组件需测试对天空背景、强光、云层等复杂空域环境的适应性。

决策与控制能力：应设计由突发风扰、通信链路暂断、动力衰减等引发的异常状态处置场景，测试部组件的自主应急策略生成能力、重规划能力与稳定性控制能力。

b） 测试指标特殊设计

效率与稳定性：需将功耗-重量比作为关键效率指标进行评估。在稳定性测试中，需重点关注在振动、高低温交变等严苛机载环境下性能指标的波动情况。

能力边界：除常规边界外，需明确界定其有效工作的海拔高度范围、最大可承受风速及抗电磁干扰能力边界。

9.3 无人车系统用部组件测评特殊要求

无人车系统强调在结构化与非结构化道路环境中，与交通规则、其他交通参与者（车辆、行人）的安全、有序交互。

a） 测试场景特殊设计

感知能力：场景库必须覆盖完备的交通要素。并专门测试对极端罕见情况的感知与识别能力。

决策与控制能力：应设计动态交互事件场景，构建长时、多任务序列，测评决策逻辑的安全性、合规性与拟人化程度。并覆盖通信断续或延迟情境，考察降级自主能力。

b） 测试指标特殊设计

可信性：功能安全与预期功能安全指标成为测评核心。需通过故障注入测试，验证部组件在失效情况下的最小风险状态实现能力。决策行为的可解释性要求极高，需能追溯其行为逻辑。

适应性：需重点测试在雨、雪、雾、眩光、夜间等不同光照与天气条件下，感知与决策性能的保持率。

9.4 无人艇系统用部组件测评特殊要求

无人艇需应对海洋环境的独特挑战，其部组件测评强调在复杂扰动、特征弱化及高腐蚀条件下的长时自主与可靠运行能力。

a） 测试场景特殊设计

感知能力：重点模拟海面波浪导致的剧烈姿态变化、低可视度（雾、雨）以及远距离弱小目标检测等场景。

导航与通信：需设计卫星导航信号拒止或欺骗场景，测试多源信息融合与自主导航的可靠性。需模拟远程、超视距通信下的高延迟、低带宽数据传输测试。

b） 测试指标特殊设计

自主性：由于远程操控延迟大，对长周期任务自主规划、动态障碍物自主避碰的能力要求高，需相应设置测评指标。

稳定性：需增加在盐雾、高湿、长周期震动环境下的耐久性与性能衰减度测试，作为长期稳定性的关键评价依据。

9.5 机器人系统用部组件测评特殊要求

机械臂、地面移动操作机器人、人形机器人等的测评核心在于在动态非结构化环境中，完成精细的肢体操作与人机近距离交互。

a） 测试场景特殊设计

感知与执行能力：需构建包含不规则地形（楼梯、碎石）、动态干扰（路过人群）、需操作物体（不同材质、形状） 的场景。对于执行部组件，需设计力控精度、抓取成功率、操作柔顺性等测试场景。

人机交互能力：须设计人类在其工作空间内突发介入的场景，测试其即时避让、安全停机等能力。

b） 测试指标特殊设计

准确性：对于执行器，定位精度、力控精度是关键准确性指标，需在负载变化时进行测试。

安全性：物理接触安全是可信性测试的重中之重，需通过模拟人机意外接触场景，测试其瞬间停止或收力的能力与响应时间。

9.6 各域无人系统部组件测评指标权重建议

权重分配应遵循4.1.1核心能力聚焦原则，针对部组件核心使命动态调整，总权重为100%。

下表给出不同无人系统测试维度权重建议，实际测评时，可根据具体部组件类型如无人机视觉感知部组件在上述权重基础上微调，并确保总权重为 100%。

表 2 可追溯性打分表示例

测评维度		无人系统				关键说明
		无人机	无人车	无人艇	机器人	
能力与性能层	能力边界	**高**(20%)	中(10%)	中(10%)	中(10%)	能源/空间约束强的平台需重点评估
	准确性	中(15%)	**高**(25%)	中(15%)	**高**(25%)	高精度定位/操作为核心的平台
	效率	**高**(20%)	中(10%)	**高**(20%)	中(10%)	续航敏感型平台
	稳定性	中(10%)	中(15%)	中(20%)	中(15%)	长期无人值守平台
智能特征	自主性	**高**(20%)	中(15%)	中(15%)	中(10%)	弱通信/高动态环境平台
	适应性	中(10%)	中(15%)	中(10%)	中(15%)	开放环境作业平台
	学习性	低(5%)	低(5%)	低(5%)	中(10%)	任务多变/人机协作平台
	可信性	中(10%)	中(5%)	中(10%)	中(15%)	高风险/人机共存平台

附录A
（资料性）
视觉感知部组件智能测评实施实例

A.1 视觉感知部组件测评设计

A.1.1 视觉感知部组件测试场景设计

A.1.1.1 视觉感知部组件测试场景（数据集）基本要求

在视觉感知部组件测评过程中，宜采用数据集对场景进行表征。数据集具体要求如下：

A.1.1 基本要求

a）测试集的数据安全应符合 GB/T 35273 的要求。

b）针对视觉感知算法的应用传感器安装位置，应采用特定的测试集进行测试。

示例1：地面感知应是地面传感器感知地面物体；

示例2：地空感知应是地面传感器感知天空物体；

示例3：空地感知应是天空传感器感知地面物体。

c）测试集应具有独立性，确保视觉感知算法在训练时没有使用过测试集中的数据。

A.1.1.2 视觉感知部组件测试场景（数据集）数量要求

测试集数量应不小于20000张。

A.1.1.3 视觉感知部组件测试场景（数据集）质量要求

a）旋转：以样本的平面视图为基准，进行0°～360°的以5°为间隔均匀旋转，样本数量不少于测试集总数量的10%。

b）视角：样本数量不少于测试集总数量的10%，测试集对视角的要求应包括：以样本的正视图为基准，从左、右两个方向以5°为间隔均匀旋转0°～30°，数量不低于测试集总数量的10%。

c）天气：测试集对天气的要求应包括：不同时间、不同天气光照强度下的数据，包括晴天（30000lx～100000lx），阴天(50lx～1000lx)，夜晚(0.2lx～10lx)，雨天(0.2lx～10lx)等不同光照下的样本数量不低于测试集的 3%。

d）多目标样本要求包括:

——多个不同类物体同时出现在同一画面中的样本；

——多个同类物体同时出现在同一画面中的样本；

——数量不低于测试集总数量的10%。

e）遮挡样本数量不少于测试集总数量的10%。

f） 噪声干扰样本数量不少于测试集总数量的10%，测试集对噪声的要求包括：

——噪声类别至少包含高斯噪声、椒盐噪声、泊松噪声等；

——噪声干扰等级至少包含五级。

g）分辨率干扰样本数量不少于测试集总数量的10%，测试集对分辨率的要求应包括：

——用于干扰场景下的测试集分辨率不低于 256×360 像素；

——分辨率干扰等级至少包含五级。

A.1.1.4 视觉感知部组件测试场景（数据集）均衡性、标注要求

a）数据集均衡性应考虑各种类别的样本数量一致程度和数据集样本分布的偏差程度。

示例：飞机的数据集包括国别等。

b）数据集标注信息应完备且准确无误。

A.1.2 视觉感知部组件测试指标选取

根据视觉感知部组件特征，选取如下的指标：

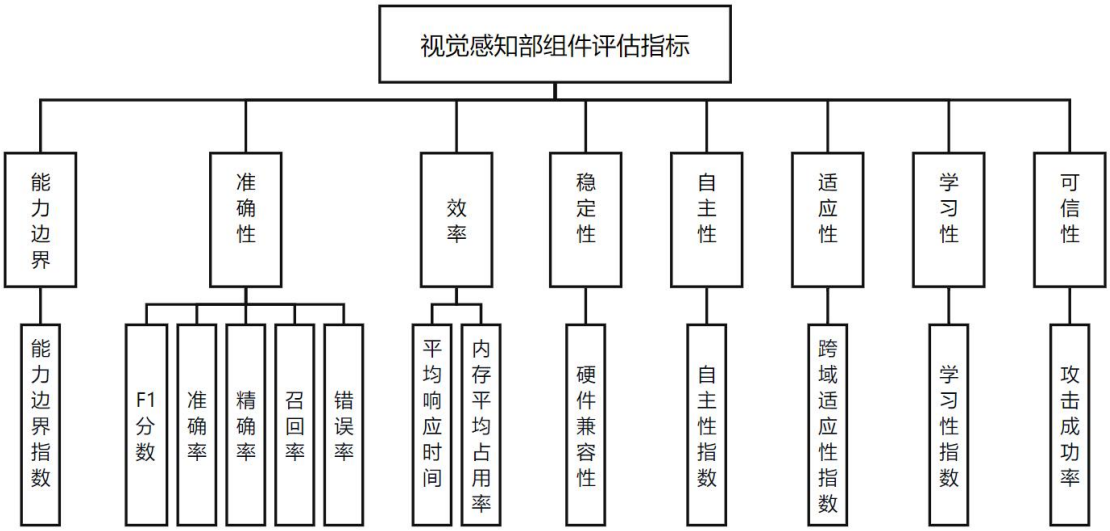

图 A.1 视觉感知部组件评估指标

A.2 视觉感知部组件测评执行

A.2.1 能力边界测试

能力边界是指视觉感知部组件在给定运行条件、输入场景及资源约束下，能够维持其预期功能与性能的最大适用范围。该指标用于描述算法在环境变化、输入扰动或任务复杂度增加时的性能稳定性与可承受极限。

a) 采用分层扰动测试方法，在基准测试集上按光照变化、视角偏移、噪声注入、遮挡比例、目标尺度变化等因素生成多层扰动数据集；

b) 在各扰动层下测试算法的主要性能指标（如 mAP、IoU、F1-score、识别精度或置信度均值等），绘制性能变化曲线；

c) 当性能下降幅度超过预设容差阈值（如基准性能下降 10%）时，判定达到能力边界点；

d) 依据扰动参数的临界值计算算法的能力边界指数。

$$B = \sum_{i=1}^{n} w_i \frac{P_{th,i}}{P_{0,i}} \quad (A.1)$$

式中:

B ——能力边界指数；

w_i —— 第i维扰动的权重系数；

$P_{th,i}$—— 性能下降至容差阈值时的扰动参数值；

$P_{0,i}$ ——正常运行条件下的基准参数值

A.2.2 准确性测试

准确性是指视觉感知部组件在执行过程中所表现出的固有性能特征，即在未进行任何特定优化或改进的情况下，模型在给定任务或数据集上的原始表现能力。

下表给出了视觉感知智能学习算法针对不同任务选取的基础性能指标示例。

表 A.1 准确性评价指标

算法任务类型	测试数据集类型	准确性指标
图像分类（二分类）	图像数据集	F1 分数、准确率、精确率、召回率、G-mean、特异度、误诊率、错误率等
图像分类（多分类）		加权平均精确率、加权平均召回率、加权平均 F1 分数、宏观平均精确率、宏观平均召回率、宏观平均 F1 分数、微观平均精确率、微观平均召回率、微观平均 F1 分数、准确率、召回率、F1 分数等
目标检测（单类/多类）		IoU、mAP、AP 明细、置信度等
目标跟踪（单类/多类）		IoU、MOTA、MOTP、IDP1、IDP、IDR、主要跟踪目标数量、主要丢失目 标数量、部分跟踪目标数量、MT、ML、PT、IDSW、碎片总数、mAP、 AP 等
语义分割		像素准确率、类别平均像素准确率、类别像素准确率、IoU、MIoU 等

A.2.3 效率测试

效率是指在给定硬件和软件条件下，视觉感知部组件进行推理操作的速度和资源消耗情况。指标包括：

a)平均响应时间:响应一个用户任务的平均时间，计算方法见公式(A.2）。

$$\bar{T} = \frac{\sum_{i=1}^{n}(T_i)}{n} \quad (A.2)$$

式中:

$\bar{T}$ ——平均响应时间；

T_i ——第 i 次测量时算法的响应时间；

n ——测得的响应次数。

b)内存平均占用率:执行一组给定的任务所需要的内存与可用内存的平均比率,计算方法见公式(A.3)。

$$OD = \frac{\sum_{i=1}^{n}\left(\frac{R_i}{RW_i}\right)}{n} \quad (A.3)$$

式中:

OD ——内存平均占用率；

R_i ——第 i 次样本处理中执行一组给定任务所占用的实际内存大小；

RW_i ——第 i 次样本处理期间可用于执行任务的内存大小；

n ——处理的样本数。

A.2.4 稳定性测试

稳定性指视觉感知部组件在特定硬件上运行时，能够有效执行其所需的功能并且不会对其他产品造成负面影响的程度,计算方法见公式 (A.4)。

$$C = C_{device}/CCN \quad (A.4)$$

式中:

C ——稳定性（硬件兼容性）;

C_{device} ——与该算法可兼容的计算处理器的数量；

CNN ——在运行环境中该算法需要兼容的计算处理器的数量

A.2.5 自主性测试

自主性是指视觉感知部组件在标注不完整或标签质量较低的弱监督条件下，能够通过内部推理、伪标签生成、自一致学习等机制，自动完成知识补全与任务执行的能力。

该指标用于衡量算法在有限人工指导下的自学习、自纠错及自决策水平。

$$A_u = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \quad (A.5)$$

分割算法

$$A_u = \frac{\text{area}（X \cap Y）}{\text{area}（X \cup Y）} \times 100\% \quad (A.6)$$

检测算法

$$A_u = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \quad (A.7)$$

式中:

A_u——自主性指数；

A.2.6 适应性测试

适应性是指视觉感知部组件在不同数据域、任务域或应用场景间迁移时，仍能保持预期感知性能和稳定输出的能力。

该指标用于衡量算法在源域与目标域存在分布差异情况下的泛化能力、特征迁移能力及模型自适应调整能力。跨域适应性用于评价算法多场景环境下的可迁移性与鲁棒性，反映其在现实复杂条件中持续发挥感知功能的能力，为算法在模型复用提供技术依据。

a）源域与目标域选择：在保持任务一致（如检测或分割）的前提下，选择具备显著分布差异的多个数据域作为测试样本；

b）基准性能测定：在源域上训练算法并记录基准性能指标（如mAP、IoU、F1-score 等）；

c）目标域迁移测试：在目标域上直接推理或经有限样本自适应后测量性能；

d）性能保持率计算：对比源域与目标域的性能变化，得到跨域性能保持率；

e）结果平均化：若存在多个目标域，计算加权平均跨域保持率以评估整体跨域适应性。

$$A = \sum_{i=1}^{n} w_i \frac{P_{t,i}}{P_{s,i}} \quad (A.8)$$

式中:

A ——跨域适应性指数；

w_i —— 第i目标域的权重系数；

$P_{t,i}$—— 算法在目标域的性能指标；

$P_{s,i}$ ——算法在源域的性能指标。

A.2.7 学习性测试

学习性是指视觉感知部组件在样本数量极少(k-shot）或训练资源受限的条件下，能够快速完成特征提取、模型自适应和任务泛化的能力。

该指标用于衡量算法在新任务、新类别或新场景下的快速学习与迁移性能。

$$S = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \quad (A.9)$$

分割算法

$$S = \frac{\text{area}（X \cap Y）}{\text{area}（X \cup Y）} \times 100\% \quad (A.10)$$

检测算法

$$S = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \quad \text{(A.11)}$$

式中:

S——学习性指数;

A.2.8 可信性测试

安全性指视觉感知部组件对对抗样本的防范能力。安全性的评估指标应包括但不限于:

攻击成功率:经过攻击方法构建的新测试数据集中,模型预测失败的样本数与总样本数之间的比率,计算方法见公式 (3)。攻击成功率越小,模型在对攻击的抵抗能力越强。

$$\mathrm{ASR} = \frac{N_{\mathrm{adv}}}{N_{\mathrm{all}}} \quad \text{(A.12)}$$

式中:

ASR —攻击成功率;

N_{all} ——样本总数;

N_{adv} ——预测失败的样本数。

A.3 视觉感知部组件能力评估

基于预处理后的数据，根据第 A.2 节定义的视觉感知部组件各项测评指标进行测试，参考第 8.2 节计算指标评估得分，将测试结果与评估得分填入表 A.2，可给出总评得分。

表 A.2 视觉感知部组件测评记录表

测评维度	测评指标	测试结果	评估得分	评估维度得分	备注
能力边界	能力边界指数				
准确性	F1 分数				准确性计算的权重分配为：基础性能 = 0.2*F1 数 + 0.2*准确率 + 0.2*精确率 +0.2*召回率 + 0.2*错误率。
	准确率				
	精确率				
	召回率				
	错误率				
效率	平均响应时间				效率指标计算的权重分配为：效率 = 0.5*平均响应时间+ 0.5*内存平均占用率。
	内存平均占用率				
稳定性	稳定性（硬件兼容性）				

测评维度	测评指标	测试结果	评估得分	评估维度得分	备注
自主性	自主性指数				
适应性	跨域适应性指数				
学习性	学习性指数				
可信性	攻击成功率				
总评得分					评估分值=0.05*能力边界+0.25*准确性+0.2*效率+0.1*稳定性+0.05*自主性+0.15*适应性+0.1*学习性+0.1*可信性
注：评估得分表示每个指标的测试结果对应的评估得分,100 分制,0 为最低分,100 为最高分,精度为小数点后两位。对于高优类指标,如 F1 分数、准确率、精确率、召回率,评估得分=测试结果×100。对于负向指标,如错误率,评估得分=(1-测试结果)×100					

参考文献

[1] GB/T 45225-2025 人工智能 深度学习算法评估
[2] GB/T 45579-2025 机器人智能化视觉评价方法及等级划分