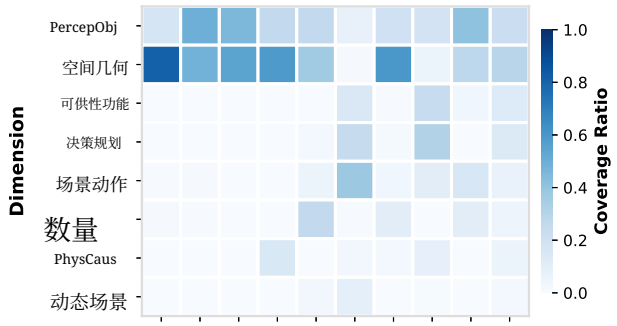
arXiv:2602.01640v1 [cs.CL] 2 Feb 2026

A2Eval：面向具身Bra的智能体与自动化评估 in

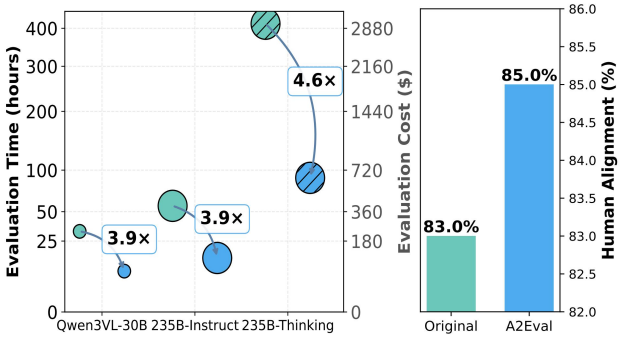
Shuai Zhang [1 2 3] **Jiayu Hu** [3] **Zijie Chen** [2] **Zeyuan Ding** [3] **Yi Zhang** [3] **Yingji Zhang** [4] **Ziyi Zhou** [3 5] **Junwei Liao** [3] **Shengjie Zhou** [6] **Yong Dai** [3] **Zhenzhong Lan** [2] **Xiaozhu Ju** [3]

摘要

当前具身视觉语言模型评估依赖静态的、由专家定义并经人工标注的基准测试，存在严重冗余和覆盖不平衡的问题。这种劳动密集型范式耗费大量计算和标注资源，推高成本并扭曲模型排名，最终阻碍迭代开发。为此，我们提出智能体自动评估（A2Eval），这是首个通过两个协作智能体自动完成基准构建与评估的智能体框架。数据智能体自主归纳能力维度，并构建平衡、精简的评估套件；评估智能体则综合生成并验证可执行评估流水线，从而实现完全自主且高保真度的评估。在 10 个基准测试和 13 个模型上进行评估后，A2Eval 将评估套件缩减了 85%，将总体计算成本降低了 77%，并实现了 4.6× 倍加速，同时保持评估质量。尤为重要的是，A2Eval 能纠正系统性排名偏差，提高与人工评估的一致性（Spearman's$\rho = 0.85$），并保持较高的排序保真度（Kendall's $\tau = 0.81$），为高保真、低成本的具身评估树立了新标准。我们的代码和数据即将公开。



（a）具身视觉语言模型基准测试中的能力覆盖度。



（b）评估成本与人工评估的一致性。

图 1：**(a)**：现有由专家定义并经人工标注的具身视觉语言模型基准测试的能力分布表明，覆盖度冗余却稀疏，导致排名失真和评估成本过高。**(b)**：A2Eval 以自动化套件取代了这种由专家定义、依赖大量标注的范式，在压缩基准测试的同时保持能力覆盖，实现了 4.6× 倍加速比并提升了与人工评估的一致性。（详情见表3 和5。）

1. 引言

评估现代具身视觉语言模型（VLM），已成为应对具身人工智能（Liu 等，2025；Cong 和 Mo，2025）快速发展的"指南针"。全面而精准的评估机制不仅是学术突破的基准测试，也是产业决策的基石（Yifan 等，2025；Zhang 等，2025a）。然而，这个"指南针"正变得成本高昂：如今，在分散的基准测试上评估单个模型，耗时已超过*3200 GPU* 小时 (见表 5)，造成沉重的资金和计算负担，阻碍了迭代研究。

[1]中国浙江大学 [2]中国西湖大学 [3]中国北京人形机器人创新中心 [4]英国曼彻斯特大学 [5]中国南方科技大学 [6]马来西亚马来西亚厦门大学。联系方式：Shuai Zhang < zhangshuai@westlake.edu.cn>。

预印本。正在评审。

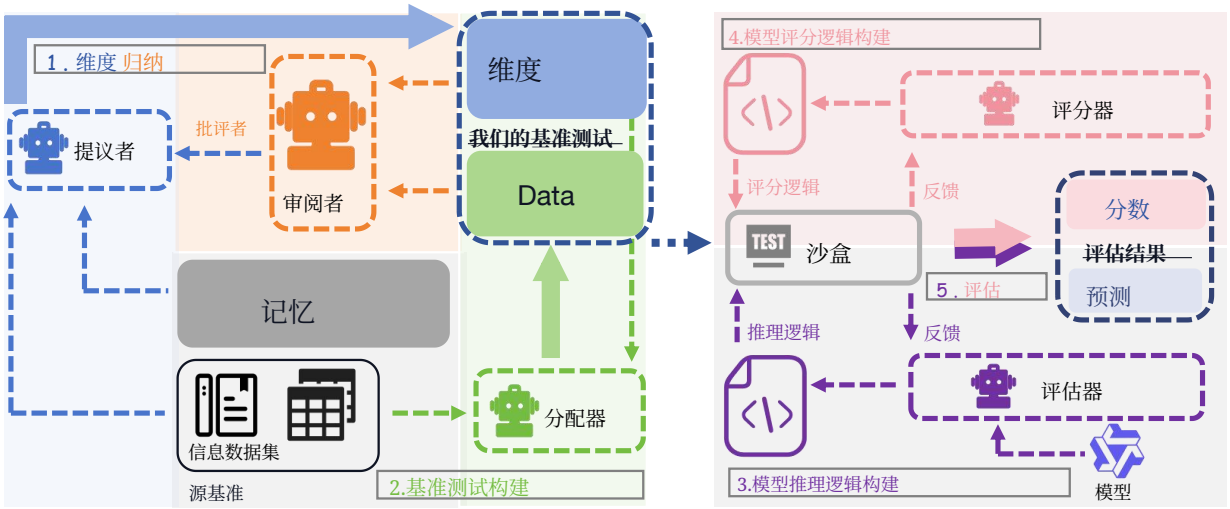


图 2：A2Eval 概览。**左**半部分展示了**数据智能体**的三种角色：提议者、审阅者和分配器，它们共同构建紧凑、均衡的基准测试。**右**半部分展示了**评估智能体**的两种角色：评估器和评分器，负责生成模型预测和分数。每种角色均以不同颜色表示。

这场危机源于失衡的评测生态系统，其根植于劳动密集型的**"专家定义 + 人工标注"**范式。如图1(a)和4所示，当前基准测试的样本相似度高达*92%*，且能力分布严重偏斜。物理因果等关键推理能力所占样本较少，而易于标注的任务则占据主导地位。这造成了一个由三种系统性问题构成的"两败俱伤"困境：(1) 覆盖不均衡与冗余：大量样本重复且任务分布偏斜，导致评估反复且不均衡，无法全面评估模型能力。(2) 排名失真：对过度代表的任务过拟合的模型尽管存在关键短板，却仍位居排行榜前列，扭曲排行榜并误导研究方向。(3) 高昂的评测成本：评估大量冗余样本，并为每个基准测试手动构建推理和评分逻辑，这两者相结合，造成了巨大的计算和人力开销，使评估周期既缓慢又昂贵，阻碍了迭代开发。

为解决这些矛盾，我们提出智能体与自动化评估（*A2Eval*），这是一种范式转变，将基准构建视为一个优化问题：在最大化能力覆盖度与多样性的同时，尽量减少冗余和人工投入。A2Eval 通过两个协作智能体构成的全自主流程，取代传统范式：

数据智能体通过自主推断能力维度，解决覆盖度失衡、冗余和排名失真问题，从而无需依赖主观的专家定义。通过自动归纳能力维度并执行多样性感知采样，它构建出平衡、紧凑的评估套件，在全面覆盖能力的同时，大幅缩减基准测试规模和冗余计算，从而纠正系统性排名偏差并显著降低评估成本。

评估智能体通过自动合成并验证可执行的评估流程，解决人工流程瓶颈。通过对推理和评分逻辑进行迭代式沙箱优化，它无需人工干预即可实现完全自主的高保真评估。

在 10 个基准测试和 13 个模型上进行的大量实验，证明了我们方法的有效性。我们的主要贡献总结如下：

(1) 首个智能体评测框架：我们提出了首个完全自动化的评估框架，使评估摆脱静态专家标签的束缚，以可扩展的自主智能体系统取代人工流程，从而应对基准测试构建与执行方面的挑战。

(2) 均衡紧凑的基准测试构建：我们的数据智能体纠正覆盖度失衡和排名失真，过滤*85%*的冗余样本，将与人类偏好的对齐度提升至*0.85*（Spearman's ρ），并将总体成本降低*77%*，同时实现*4.6*× 倍的评估吞吐量提升。

(3) 自主可执行评估流水线：我们的评估智能体会自动构建经过验证的推理和评分逻辑，在实现完全自主评估、无需人工干预的同时，达到*96.9%*的保真度。

2. 相关工作

2.1. 具身推理基准

近期的基准测试评估了广泛的具身推理能力。Where2Place（Yuan 等，2025）和 VSI-Bench（Yang 等人，2025a）等任务着眼于放置和空间理解；RefSpatial（Zhou 等，2025）、RoboSpatial（Song 等，2025）、EmbSpatial（Du 等，2024）和 OmniSpatial（Jia 等，2025）则将空间推理扩展到接地和操作。ERQA（Team 等，2025b）和 COSMOS（Azzolini 等，2025）侧重于物理推理和因果推理，而 BLINK（Fu 等，2024）和 EgoSchema（Mangalam 等人，2023）则评估多模态感知和自我中心理解。近期的视觉语言模型（例如 Qwen3-VL（Bai 等，2025a）、InternVL3.5（Wang 等，2025））以及具身专用系统（例如 Pelican-VL（Zhang 等，2025a;b）、RoboBrain（Team 等，2025a））越来越多地采用此类基准测试，同时也出现了 EmbodiedBench（Yang 等，2025b）和 EAI（Li 等，2024）等系统级框架。然而，基准测试的迅速增多带来了评估挑战，包括大量冗余（Zhang 等，2025d）和高昂的计算成本，尤其是对于基于长视频的基准测试（Mangalam 等人，2023；Yang 等人，2025a）。

2.2. 智能体评估与基准

并行研究探索了自动化基准构建。Code2Bench（Zhang 等，2025c）与 OK-Bench（Li 等，2025）提供了领域专用流程，用于生成具备污染控制与可复现性协议的新基准。然而，这些方法侧重于创建全新的评估集，而非整合现有异构基准，也未应对冗余与计算成本。我们的工作与先前工作的区别在于引入了（1）跨领域均衡采样的自动能力归纳，（2）面向冗余的基准整合，以及（3）统一的端到端数据与评估流程。

3. 方法

我们提出智能体自动评估（A2Eval），以应对具身领域中基准测试冗余、覆盖度碎片化和评估成本高昂等挑战。如图2所示，A2Eval 是一个双智能体框架，可自动完成从基准测试构建到模型评估的整个生命周期。该框架由数据智能体（第3.1节）和评估智能体组成。数据智能体通过专门的智能体协作（**提议者、审阅者和分配器**）发掘统一的能力维度，并通过分配器进行维度分配和多样性感知采样，构建紧凑、均衡的基准测试；评估智能体则（第3.2节）通过迭代代码合成生成可执行的评估流水线。有关详细的算法流程和提示设计，请参阅附录C。

3.1. 数据智能体

给定现有的基准信息池 $\mathcal{B} = \{B_i\}_{i=1}^{N_B}$及其示例集 $\mathcal{E} = \bigcup_{i=1}^{N_B} E_i$，其中 N_B 为基准测试的数量，数据智能体会为目标领域构建一个具备能力感知、低冗余且平衡的基准测试。其核心创新在于将基准测试构建视为一个两阶段的智能体工作流：**(i) 维度归纳**，由多个专用智能体协同发现统一的能力分类体系；**(ii)基准测试构建**，由分配器执行维度分配和多样性感知采样，从而生成一个平衡、无冗余且保持全面语义覆盖的测试集。

维度归纳。 维度归纳通过三个角色各异、推理能力不同的专用智能体进行迭代协作：生成候选能力维度的**提议者** $\mathcal{A}_p$；从覆盖度、独立性和平衡性角度批评提案的**审阅者** $\mathcal{A}_r$；以及通过示例分配和多样性感知采样报告维度平衡情况的**分配器** $\mathcal{A}_a$。这些智能体通过共享记忆 M 进行交流，其中记录了历史提案和批评意见，从而支持它们协同推理，确定最优维度集。

在第 i 次迭代中，提议者通过推理基准测试信息 $\mathcal{B}$、审阅者的批评意见 C_i 和历史记忆 M，生成候选维度集 D_{i+1}：

$$D_{i+1} = \mathcal{A}_p(\mathcal{B}, C_i, M). \tag{1}$$

提议者 根据来自 C_i 的具体反馈完善维度，同时结合从 M 中获得的经验。

审阅者 根据三项标准评估 D_{i+1}：(a) 维度之间的概念冗余；(b) 相对于领域要求的覆盖全面性；(c) 维度平衡。批评意见 C_{i+1} 由以下方式生成：

$$C_{i+1} = \mathcal{A}_r(D_{i+1}, S^{(r)}). \tag{2}$$

其中， $S^{(r)}$ 包含来自分配器的平衡统计数据。每次迭代后，记忆都会更新为 $M \leftarrow M \cup \{(D_{i+1},\ C_{i+1})\}$。

提议者和审阅者不断迭代，直至维度趋于稳定（$D_{i+1} = D_i$），并生成候选方案 $D^{(r+1)}$。为验证平衡性，分配器会通过实际示例分配和抽样来测试 $D^{(r+1)}$，从而得出平衡统计数据$S^{(r+1)}$。审阅者会将这些实证证据

维度	定义
感知 & 物体定位（PercepObj）	定位、识别和跟踪物体，例如选择坐标或掩码、检测边界框以及选择抓取点。
场景 & 动作 理解（场景动作）	解读场景、动作、状态和姿态，例如识别环境、人类或智能体活动、物体状态以及身体或姿态线索。
空间与 几何推理（空间几何）	理解空间关系、参照系和拓扑结构，包括位置、朝向、包含关系和邻接关系。
定量与数值推理（数量）	确定或比较数值属性，包括物品计数、估算连续量以及判断相对大小。
可供性与功能推理（可供性功能）	评估物体可供哪些动作，例如抓握、打开或倾倒，以及工具使用，并评估在约束条件下进行操作的可行性。
物理 & 因果 推理 (PhysCaus)	对物理可行性和稳定性进行推理，例如判断适配性、支撑情况，平衡状态，并预测碰撞或其他物理相互作用。
决策 & Task 规划 (决策规划)	跨步骤进行规划和决策，包括选择后续动作、应用条件逻辑并将任务分解为子目标。
动态 场景 推理（动态场景）	理解物体与整个场景如何变化、交互并保持一致-随时间推移，例如事件顺序、状态和空间关系。

表1：数据智能体归纳的能力维度（左：全称与缩写）及详细定义（右）。

纳入下一轮评议，进而触发维度优化（例如，合并代表性不足的维度）。这一智能体驱动的验证循环会持续进行，直至生成稳定的维度集 D。

基准测试构建。给定已验证的维度集合 D，**分配器**通过两阶段流程构建基准测试：先进行维度分配，再进行多样性感知采样。这将维度分类体系转化为实际测试集，在消除冗余的同时保持全面的能力覆盖。

维度分配。分配器采用由 N_v 个投票智能体 $\{\mathcal{A}_a^1, \ldots, \mathcal{A}_a^{N_v}\}$ 组成的集成，它们独立推理能力需求。每个投票智能体 $\mathcal{A}_a^k$ 分析示例 $e_j \in \mathcal{E}$，并基于自身解释生成维度分配D_{jk}。最终分配 D_j由多数共识产生：

$$D_j = \text{MajorityVote}\big(\{D_{jk}\}_{k=1}^{N_v}\big). \tag{3}$$

逐维度示例池聚合为

$$E_{d_k} = \{e_j \in \mathcal{E} : d_k \in D_j\} \tag{4}$$

对每个 $d_k \in D$。

多样性感知采样。完成维度分配后，分配器通过基于聚类的选择执行多样性感知采样，在保留语义多样性的同时消除维度内冗余。对于每个维度 d_k，将样本 $e \in E_{d_k}$ 嵌入为$x = \text{Encoder}(e)$，映射到语义空间。如果 $|E_{d_k}| < K$，则保留所有样本；否则，选择距离每个 K 聚类中心 μ_j最近的样本：

$$E_{d_k}^{\text{s}} = \Big\{ \arg\min_{x \in C_j} \|x - \mu_j\| \Big\}_{j=1}^{K}. \tag{5}$$

维度	源集合	保留集合
PercepObj	6,218 (25.4%)	500（13.2%）
场景动作	1,864（7.6%）	500（13.2%）
空间几何	10,565（43.1%）	500（13.2%）
数量	3,185（13.0%）	500（13.2%）
可供性功能	773（3.2%）	500 (13.2%)
物理因果	366 (1.5%)	366 (9.7%)
决策规划	1,104 (4.5%)	500 (13.2%)
动态场景	425 (1.7%)	425 (11.2%)
其他	19 (0.0%)	0 (0.0%)
总计	**24,519**	**3,781**

表 2：固定-K 聚类采样前后各能力维度的样本数量。“源集合”表示由数据智能体标注的样本；“保留集合”表示经多样性感知采样后保留的样本。百分比显示在括号中。

这种基于质心的选择确保所保留的样本$E_{d_k}^{\text{s}}$ 能够稀疏而有策略地覆盖语义流形（如图3所示），在不牺牲表征完整性的情况下去除冗余。通过统一应用K 于各维度，实现了跨维度平衡，同时每个维度保持最大的内部多样性。最终基准测试的构建方式为：

$$\mathcal{E}^{\text{s}} = \bigcup_{k=1}^{N_d} E_{d_k}^{\text{s}}. \tag{6}$$

以及平衡统计数据

$$S = \{|E_{d_k}^{\text{s}}| : d_k \in D\}. \tag{7}$$

3.2. 评估智能体

评估智能体会对数据智能体生成的基准测试执行自动化评估。它包含两个特殊的——

化角色：(i) 构建模型推理逻辑的**评估器**，以及 (ii) 构建性能评估逻辑的**评分器**。这些角色与沙箱执行器协作，迭代生成并验证完整的评估流程。

模型推理逻辑构建。 **评估器**$\mathcal{A}_e$ 合成推理逻辑，将目标模型 M 与从 $\mathcal{E}^s$采样的基准测试样例 $\mathcal{T}$相结合。在迭代 i时，评估器会生成可执行的推理代码 F_i^e，实现以下功能：加载测试输入、调用 M并生成模型预测。

沙箱执行器 $\mathcal{X}$ 使用测试用例 $\mathcal{T}$在模型 M 上运行 F_i^e。如果执行成功，则完成经验证的逻辑 $F^e = F_i^e$。否则，$\mathcal{X}$ 会向评估器 E_i返回诊断反馈，以更新推理逻辑：

$$F_{i+1}^e = \mathcal{A}_e(M, \mathcal{T}, F_i^e, E_i). \tag{8}$$

这一合成与验证循环会持续迭代，直到 F^e 能够可靠执行或达到预设的迭代预算。最终产物 F^e 构成用于系统化模型评估的经验证推理逻辑。

模型评分逻辑构建。 **</style}给定经验证的推理逻辑以及模型在测试用例 上生成的模型预测（，$\mathcal{T}$），评分器** $\mathcal{A}_s$ 会合成评估逻辑，以评估模型表现。在迭代 j 时，评分器会根据测试用例 $\mathcal{T}$和模型输出 $\mathcal{P}_\mathcal{T}$生成可执行的评分代码 F_j^s:

沙箱执行器 $\mathcal{X}$ 在测试输出 $\mathcal{P}_\mathcal{T}$上运行 F_j^s。如果执行成功并生成有效指标，则逻辑$F^s = F_j^s$最终确定。否则，$\mathcal{X}$会向评分器返回诊断反馈 E_j，触发迭代优化：

$$F_{j+1}^s = \mathcal{A}_s(\mathcal{T}, \mathcal{P}_\mathcal{T}, F_j^s, E_j). \tag{9}$$

综合与验证循环会持续进行，直到 F^s 能够可靠执行或达到迭代预算。由此生成的产物F^s 构成了可复现的评分逻辑，可用于全面的模型评估。

基准评估执行。 推理逻辑 F^e和评分逻辑 F^s均通过验证后，便会在构建的基准测试 $\mathcal{E}^s$上运行完整的评估流程。评估器应用 F^e，为所有测试用例生成模型预测：

$$\mathcal{P} = F^e(M, \mathcal{E}^s), \tag{10}$$

评分器应用 F^s 计算最终评估指标：

$$\text{Results} = F^s(\mathcal{E}^s, \mathcal{P}). \tag{11}$$

这将为目标模型 M 在 D 的所有能力维度上生成全面的评估结果。

模型	$\widetilde{\mathcal{B}}$	$\mathcal{B}$	$\mathcal{H}$
Qwen3-VL-235B-A22B-Thinking	65.97	61.26	87.9
Internvl-3.5-241B-A28B	65.68	56.76	63.6
GPT-5-20250807-Mini	65.52	55.16	79.4
Qwen3-VL-30B-A3B-Thinking	62.58	54.14	71.2
Internvl-3.5-38B	62.37	55.01	64.5
Qwen3-VL-235B-A22B-Instruct	61.88	55.39	77.4
Internvl-3.5-30B-A3B	61.79	51.38	55.3
Qwen3-VL-30B-A3B-Instruct	61.47	48.36	62.5
Internvl-3.5-8B	55.98	48.84	61.5
Qwen2.5-VL-72B-Instruct	54.76	51.07	61.3
Qwen2.5-VL-32B-Instruct	50.77	47.24	58.5
Qwen2.5-VL-7B-Instruct	48.22	44.26	55.5
Qwen2.5-VL-3B-Instruct	39.39	37.62	50.4
与人类的排名相关性（$\mathcal{H}$）			
斯皮尔曼的 ρ	**0.85**	0.83	
肯德尔的 τ	**0.72**	0.64	
~我们（$\mathcal{B}$）与来源（$\mathcal{B}$）之间的排名相关性			
斯皮尔曼的 ρ		0.94	
肯德尔的 τ		0.81	

~表3：模型在我们构建的基准测试 $\mathcal{B}$、原始基准测试 $\mathcal{B}$以及人工评估 $\mathcal{H}$上的得分。底部报告了与人工评估之间以及两个基准测试之间的秩相关。

4. 实验

我们在具身视觉—语言推理领域评估 A2Eval。在这一领域，模型必须基于视觉输入，综合完成感知、空间推理、物理理解和多步决策。我们验证了数据智能体能够生成语义上有意义的能力维度和均衡的基准测试，并与人工判断排名相一致。此外，我们强调，该方法显著提高了评估效率，而评估智能体在自动推理和评分方面准确率很高。综合来看，这些结果表明，A2Eval 能够实现完全自动化、高效且可信的能力感知评估。

4.1. 实验设置

基准测试。我们聚焦于具身视觉语言推理领域，采用了十个成熟的基准测试：COSMOS（Azzolini 等，2025年）、ERQA（Team 等，2025b）、Where2Place（Yuan 等，2025年）、VSI-Bench（Yang 等人，2025a）、OmniSpatial（Jia 等，2025年）、EgoSchema（Mangalam 等人，2023年）、BLINK（Fu 等，2024年）、RefSpatial（Zhou 等，2025年）、RoboSpatial（Song 等，2025年）和 EmbSpatial（Du 等，2024年）。这些数据集既为数据智能体构建基准测试提供依据，也作为参考语料库，用于分析冗余、覆盖度和评估效率。关于这些基准测试及其冗余样本的详细信息

见附录A。

模型。 我们评估了多种不同规模和架构的视觉语言模型，包括若干 Qwen2.5-VL 模型（Bai 等，2025b）、多个 InternVL-3.5 变体（Wang 等，2025）、最新的 Qwen3-VL 系列（Bai 等，2025a），以及闭源模型 GPT-5 Mini。所有评估均使用 lmms-eval 框架（Zhang 等，2024年）以批量大小 1 进行。

评估设置。 默认情况下，大规模模型（例如 Qwen3-VL-235B-A22B-Thinking）使用 vLLM 后端部署，而较小的模型（例如 Qwen2.5-VL-7B-Instruct）则通过 Accelerate 使用 PyTorch 后端，并采用 8 路数据并行。所有模型均采用贪心解码，以确保输出具有确定性。视频输入统一以 2 FPS 采样 32 帧，最大生成长度设为 4096 个词元。更多详情见附录B。

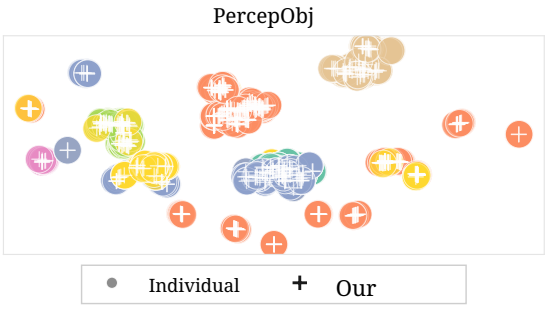
A2Eval 设置。 对于智能体工作流，我们使用 GPT-4o 作为分配器，而提议者、审阅者和评估智能体则由 Gemini 3 Pro 实现。在分配器为数据智能体进行多样性感知采样时，通过拼接视觉和文本嵌入对示例进行编码。文本内容使用句子表征模型进行嵌入[1]，视觉表征则使用 CLIP 提取（Radford 等人，2021）。对于视频输入，会处理均匀采样的 32 帧，并通过对这些帧的帧级 CLIP 特征取平均值，获得最终的视觉嵌入。所有维度的簇数 K 均设为 500，以平衡语义覆盖和评估效率。

4.2. 总体结果

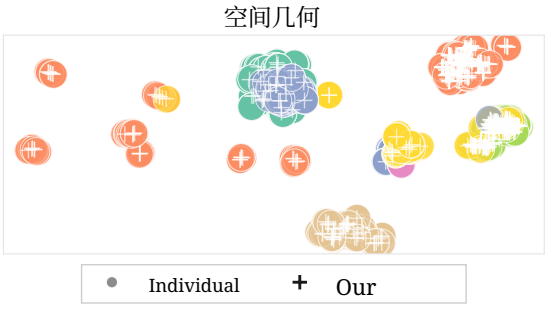
统一维度。 我们的数据智能体自动归纳出八个用于具身视觉—语言推理的能力维度。表1列出了每个归纳出的维度，并附有简洁的缩写和智能体生成的完整定义。

均衡基准。 桌子2 报告分配到各个维度并在分配器的多样性感知采样后保留的样本数量。由于规模可忽略不计，省略“其他”类别。原始池包含 24,519 个样本，而采样基准包含 3,781 个样本。

覆盖度与多样性。图1（a）展示了原始基准测试在各维度上的分布情况。基准测试强调了不同的能力，但各维度之间存在大量重叠。这表明，汇总现有基准测试虽能实现广泛覆盖，却也会造成冗余，因此应进行能力感知的整合，而非简单合并。图3 展示了两个代表性维度 PercepObj 和空间几何的样本嵌入在 UMAP（McInnes 等人，2018）投影中的分布（其他维度见图6）。我们选出的子集（以“+”符号标记）稀疏但有策略地覆盖了整个语义流形，表明我们的数据智能体在消除维度内冗余的同时，保留了全面的能力覆盖。



（a）：感知物体定位模块。



（b）：空间几何推理模块。

图 3：样本嵌入的 UMAP 可视化，表明尽管稀疏度较高，我们的统一子集仍保持了全面的语义覆盖。

4.3. 诱导能力维度的质量

我们考察**数据智能体** 诱导的能力维度是否（i）作为具身视频语言推理的表征，在概念上定义明确；以及（ii）在实践中能否可靠地分配给各个样本。

维度构建质量。我们从两个方面评估诱导出的能力空间：全面性（维度是否覆盖具身大脑领域所需的核心能力）和独立性 （维度是否捕捉到不同的推理信号）。

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

全面性。这八个维度系统地分解了具身智能体的认知需求，涵盖了从感觉输入到动作生成的处理层级。在感知层面，PercepObj有助于实体接地，而SceneAct则解读语义语境和智能体行为。转至推理层面，SpatGeo用于解析几何关系，PhysCaus则确保遵循物理定律并维持稳定性；与此同时，Affd-Func评估操作可行性，QuantNum计算数值属性。最后，在执行层面，DecPlan统筹动作序列编排和目标分解，DynScene则负责监测时间演化和状态转移。这种层级组织方式与具身大脑系统的功能架构相一致；在该架构中，视觉语言模型必须连接感知、推理和规划，才能实现与现实世界的交互。表2中的分类统计验证了这种覆盖度：在24,519个样本中，仅有19个被归类为“其他”，这表明该分类体系无需借助辅助类别便能涵盖具身领域。

独立性。维度依据推理机制而非输入模态或任务表象加以区分。例如，SpatGeo（拓扑关系）和PercepObj（实体定位）都处理视觉输入，但涉及不同的认知操作；前者需要关系抽象，后者则需要感知接地。同样，PhysCaus（被动动力学）和AffdFunc（面向行动的可供性）都涉及物理理解，但二者的区别在于，智能体推理的是自然法则还是干预可行性。

4.4. 基准有效性与合理性

除了考察覆盖度和冗余削减之外，我们还检验所构建的基准测试是否：（i）产生与人类判断一致的维度分配；（ii）在完整的多基准设置下保持排序保真度；以及（iii）得出更合理、具备能力感知的排序，更准确地反映人类感知到的具身能力。

维度分配上的人机一致性。为评估自动维度分配与人类理解之间的一致性，我们与 5 位领域专家开展了一项一致性研究（附录 D）。标注员在盲审条件下审查模型提出的标签，并通过多数投票确定最终结果。我们发现，自动标注与经人工验证的标注具有很强的一致性，Cohen’s κ 平均值为 0.78，标注者间一致性为 0.80。这些结果证实，我们的基准测试所产生的维度分配与人类判断高度一致。

压缩后的排序保真度。我们将~模型在我们构建的基准测试 $\mathcal{B}$ 上的排名与所有原始基准测试上的排名 $\mathcal{B}$进行比较。桌子3 表明，我们的方法实现了较强的一致性（$\rho = 0.94$，$\tau = 0.81$），这说明去除冗余大幅减少了评估量（从 24,519 个样本降至 3,781 个样本），同时保留了整体比较结构。

通过与人工评估的一致性实现能力感知排序。尽管整体排序保持稳定，重新平衡仍纠正了受数据集偏差或过度采样能力影响的模型相对次序。为验证这些变化反映的是评估质量的提升，而非随机波动，我们基于 8 个逼真的具身场景构建了以人为中心的排序（附录 E），涵盖家庭、杂货店、实验室、仓库和临床环境。在每个场景中，标注员观看真实世界视频，并在匿名条件下评估模型逐步制定的具身行动计划。每项输出都依据八项与能力相对应、反映所诱导维度的标准进行评分，采用 1–5 分的李克特量表。由此得出的人类感知排序作为独立参照，用于评估能力感知重新平衡与人类判断之间的一致性。

表3 显示，我们的基准测试与人类排名高度一致（$\rho = 0.85$，$\tau = 0.72$），优于原始基准测试与人类排名的相关性（$\rho = 0.83$，τ =0.64）。模型排名的提升表明其跨维度推理能力更强，也更符合人类偏好。这说明我们的基准测试能生成更合理、更能体现能力的排序，反映真正的具身能力，而非数据集伪影。

Qwen3-VL-30B-A3B-Instruct 与 Qwen2.5-VL-72B-Instruct 的案例研究。如表 3和表7所示，原始 $\mathcal{B}$ 将 Qwen2.5-VL-72B-Instruct 排在~更高位置（51.07 对 48.36），但 $\mathcal{B}$ 给出了相反的排序（61.47 对 54.76），与人类偏好一致（62.5 对 61.3）。排名反转源于对维度覆盖不平衡的纠正。在 $\mathcal{B}$中，PhysCaus（1.5%）和 DecPlan（4.5%）等关键推理能力测试不足，而 SpatGeo 占据主导地位（43.1%）。重新平衡后（PhysCaus：9.7%，DecPlan：13.2%，SpatGeo：13.2%），差距显著显现：Qwen3-VL-30B-A3B-Instruct 在 PhysCaus 上领先 21.5 分（67.10 对 45.60），在 DecPlan 上领先 11.1 分（66.30 对 55.16）；由于权重较低，这些短板几乎不会影响 $\mathcal{B}$ 中的汇总分数。重新平衡后的排名如今体现了稳定的跨维度优势，也与人类判断一致。

采样策略的消融研究。表4展示了我们对原始基准测试进行智能体采样时，各个组件对评估质量的贡献。**按来源比例**采样保留了原始基准测试的分布，获得了最高的 $\mathcal{B}$ 相关性（$\rho = 0.98$）

策略	智能体来源	智能体人工
按来源比例	**0.98**	0.81
仅维度感知	0.92	0.82
我们的方法（维度 +多样性）	0.94	**0.85**

表 4：数据智能体采样策略的消融研究。Ours (Dim. +Div.) 表示我们采用的具备维度感知和多样性感知能力的智能体采样方法。Spearman's ρ 衡量与原始基准测试分数（Agn. Source）和人工评估（Agn. Human）的排名相关性。

但与人工评估的一致性最低（$\rho = 0.81$），这表明单纯照搬原始结构会保留其偏差。**仅维度感知** 通过多智能体投票分配维度，构建智能体维度，同时在每个维度内采用均匀采样，将与人工评估的一致性提升至 $\rho = 0.82$，这是通过平衡各维度的样本实现的。**我们的方法**（维度感知 + 多样性感知）进一步在每个维度内采用多样性感知采样，在保持合理的 $\mathcal{B}$ 一致性（$\rho = 0.94$）的同时，实现了最高的与人工评估的一致性（$\rho = 0.85$）。这两个组成部分都不可或缺：通过多智能体投票实现的智能体维度感知可平衡各维度的样本，而通过基于聚类的采样实现的多样性感知则能消除维度内冗余，同时不牺牲语义覆盖度。

4.5. 计算效率

我们的基准测试在全面评估能力的同时，大幅降低了评估成本。表 5 报告了 7B–241B 模型的实际评估耗时。该基准测试在所有模型上均实现了稳定的 3.4× 至 4.6× 倍加速。对于 Qwen3-VL-235B-A22B-Thinking，评估时间从 412.9 小时缩短至 89.4 小时（4.6×）；而 InternVL-3.5-241B-A28B 的评估时间则从 80.0 小时缩短至 18.9 小时（4.2×）。这一效率提升对前沿规模模型至关重要，因为对其进行完整的多基准测试评估会变得成本高昂、难以承受。即使是 Qwen2.5-VL-7B-Instruct 这样的小模型，也实现了 3.4× 倍加速（从 2.4 小时缩短至 0.7 小时），表明不同规模的模型均能持续受益。

4.6. 评估智能体忠实度

除效率之外，我们通过将评估智能体的分数与参考实现进行比较，评估其正确性与忠实度。

评估协议。 我们通过计算智能体生成的评估分数与参考评估分数之间的相对差异来衡量整体保真度：保真度 $= 1 - |\text{分数}_{\text{agent}} - \text{分数}_{\text{ref}}| / \text{分数}_{\text{ref}}$，其中分数$_{\text{agent}}$ 表示完整的智能体生成评估流程所产生的最终分数（包括

模型	来源	我们的	加速比
Qwen3-VL-235B-A22B-Thinking	412.9	89.4	4.6×
Qwen3-VL-235B-A22B-Instruct	56.0	14.5	3.9×
InternVL-3.5-241B-A28B	80.0	18.9	4.2×
Qwen3-VL-30B-Instruct	32.3	8.3	3.9×
Qwen2.5-VL-72B-Instruct	11.2	3.0	3.7×
InternVL-3.5-8B	3.8	0.9	4.2×
Qwen2.5-VL-7B-Instruct	2.4	0.7	3.4×

表5：使用8个GPU时的评估时间（小时）对比。S.D.表示我们的方法相较于原始基准测试的加速比。我们的方法在所有模型上均实现了稳定的3.4×至4.6×倍加速比。

推理逻辑 F^f 和评分逻辑 F^s），而分数$_{\text{ref}}$表示参考实现得出的分数。我们在全部八个能力维度上评估了三个模型（Qwen2.5-VL-7B-Instruct、Qwen3-VL-8B-Instruct、InternVL-3.5-8B）。

维度	Q2.5-7B	Q3-8B	IVL3.5-8B	平均值
PercepObj	92.0	93.7	94.1	93.3
场景动作	97.2	98.5	98.6	98.1
空间几何	95.7	94.4	91.0	93.7
数量	90.5	97.4	93.8	93.9
可供性功能	99.2	96.5	100.0	98.6
PhysCaus	98.4	96.4	98.1	97.6
决策规划	96.4	95.5	94.5	95.5
动态场景	94.1	96.6	99.2	96.6
最终分数	97.2	96.2	97.4	**96.9**

表 6：评估智能体在所有能力维度上的总体保真度。数值表示智能体生成的评估分数与参考分数之间的保真度。该智能体的平均保真度达到 96.9%，展现了可靠的端到端评估能力。

保真度结果。表6 展示了所有能力维度和模型的总体评估保真度。评估智能体的平均保真度达到 96.9%，表明其能够高度可靠地生成评估流水线。此端到端评估验证了从推理到评分的完整流水线所产生的结果与参考实现高度一致。推理和评分保真度各组成部分的详细分析见附录F。

5. 结论

在这项工作中，我们提出了智能体自动评估（A2Eval），这是首个利用两个协作智能体自动构建基准测试的智能体框架。数据智能体自主归纳能力维度，以构建平衡的评估套件；评估智能体则自动合成并验证可执行评估

流水线。大量实验表明，A2Eval 能够缓解基准覆盖不均衡、评估排名失真和计算成本问题。

影响声明

本文旨在推动机器学习领域的发展，具体聚焦于具身视觉语言模型评估。我们提出的框架 A2Eval 能够实现更高效、高保真且公平的评估，降低计算成本并减轻系统性的排名偏差。虽然我们未预见到直接的伦理风险或负面的社会后果，但计算资源的减少可能通过降低能源消耗而对环境产生积极影响。总体而言，我们的工作旨在通过提升评估质量与效率来支撑研究群体，且不会引入任何已知的有害影响。

参考文献

Azzolini, A.、Bai, J.、Brandon, H.、Cao, J.、Chattopadhyay, P.、Chen, H.、Chu, J.、Cui, Y.、Diamond, J.、Ding, Y. 等。Cosmos-reason1：从物理常识到具身推理。*arXiv* 预印本 *arXiv:2503.15558*，2025。

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl 技术报告。*arXiv* 预印本 *arXiv:2511.21631*，2025a。Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025b。

Cong, Y. 和 Mo, H. 基于多模态模型的机器人具身智能综述：任务、模型与系统方案。国际智能系统杂志，2025(1):5124400, 2025年。

Du, M.、Wu, B.、Li, Z.、Huang, X.-J. 和 Wei, Z. EmbSpatial-Bench：使用大型视觉语言模型对具身任务的空间理解进行基准测试。在第 *62* 届计算语言学协会年会论文集（第二卷：短论文）中，第 346–355 页，2024年。

Fu, X.、Hu, Y.、Li, B.、Feng, Y.、Wang, H.、Lin, X.、Roth, D.、Smith, N. A.、Ma, W.-C. 和 Krishna, R. BLINK：多模态大语言模型能够看见，但无法感知。在欧洲计算机视觉会议中，第 148–166 页。Springer，2024年。

Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H. 和 Yi, L. OmniSpatial：面向视觉语言模型的综合性空间推理基准。*arXiv* 预印本 *arXiv:2506.03135*，2025。

Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E. L., Zhang, R. 等。具身智能体接口：对用于具身决策的 LLM 进行基准测试。神经信息处理系统进展，37:100428–100534, 2024年。

Li, Y., Xu, T., Tang, K., Livescu, K., McAllester, D., and Zhou, J. Okbench：通过完全自动化、按需提供的开放知识基准测试，让大语言模型评测普惠大众。*arXiv*预印本 *arXiv:2511.08598*，2025。

Liu, H., Guo, D., and Cangelosi, A. 具身智能：形态学、行动、感知与学习的协同作用。*ACM*计算调查，57(7):1–36，2025。

Mangalam, K., Akshulakov, R., and Malik, J. EgoSchema：用于超长视频语言理解的诊断基准。神经信息处理系统进展，36:46212–46244，2023。

McInnes, L., Healy, J., and Melville, J. UMAP：用于降维的统一流形逼近与投影。*arXiv*预印本 *arXiv:1802.03426*，2018。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等. 从自然语言监督中学习可迁移视觉模型. 发表于国际机器学习会议，第 8748–8763 页。PMLR，2021。

Song, C. H., Blukis, V., Tremblay, J., Tyree, S., Su, Y. 和 Birchfield, S. RoboSpatial：为机器人学中的二维与三维视觉语言模型教授空间理解. 发表于计算机视觉与模式识别会议论文集，第 15768–15780 页，2025。

Team, B. R., Cao, M., Tan, H., Ji, Y., Chen, X., Lin, M., Li, Z., Cao, Z., Wang, P., Zhou, E., 等. Robobrain 2.0 技术报告。*arXiv* 预印本 *arXiv:2507.02029*，2025a。

Team, G. R., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., 等。Gemini机器人：将人工智能带入物理世界。*arXiv*预印本 *arXiv:2503.20020*，2025b。

Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., 等。Internvl3. 5：在通用性、推理和效率方面推进开源多模态模型。*arXiv*预印本 *arXiv:2508.18265*，2025。

Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L. 和 Xie, S. 空间思维：多模态大语言模型如何观察、记忆和回忆空间。载于计算机视觉与模式识别会议论文集，第 10632–10643 页，2025a。

Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T. V., Movahedi, M., Li, M. 等。EmbodiedBench：面向视觉驱动的具身智能体，对多模态大语言模型进行全面基准测试。*arXiv* 预印本 *arXiv:2502.09560*，2025b。

Yifan, C., Wei, M., Wang, X., Liu, Y., Wang, J., Song, H., Ma, L., Di, D., Sun, C., Liu, K. 等。具身人工智能：从感知智能到行为智能的演进综述。*SmartBot*，1(3):e70003，2025。

Yuan, W.、Duan, J.、Blukis, V.、Pumacay, W.、Krishna, R.、Murali, A.、Mousavian, A. 和 Fox, D. Robopoint：用于机器人学中空间可供性预测的视觉语言模型。载于机器人学习会议，第 4005–4020 页。PMLR，2025。

Zhang, K.、Li, B.、Zhang, P.、Pu, F.、Cahyono, J. A.、Hu, K.、Liu, S.、Zhang, Y.、Yang, J.、Li, C. 和 Liu, Z. Lmms-eval：对大型多模态模型评估的现实检验，2024年。网址：https://arxiv.org/abs/2407.12772。

Zhang, Y.、Liu, C.、Ren, X.、Ni, H.、Zhang, S.、Ding, Z.、Hu, J.、Shan, H.、Niu, Z.、Liu, Z. 等。Pelican-vl 1.0：面向具身智能的基础大脑模型。*arXiv* 预印本 *arXiv:2511.00108*，2025a。

Zhang, Y.、Liu, C.、Ren, X.、Ni, H.、Zhang, Y.、Zhang, S.、Ding, Z.、Hu, J.、Shan, H.、Qi, J. 等。通过刻意练习策略优化连接视觉语言模型与具身智能。*arXiv* 预印本 *arXiv:2511.16602*，2025b。

Zhang, Z.、Liu, R.、Liu, A.、Liu, X.、Gao, X. 和 S un, H.。用于评估大语言模型现实世界代码能力的动态基准构建。*arXiv* 预印本 *arXiv:2508.07180*，2025c。

Zhang, Z.、Zhao, X.、Fang, X.、Li, C.、Liu, X.、Min, X.、Duan, H.、Chen, K. 和 Zhai, G.。多模态大语言模型基准的冗余原则。*arXiv* 预印本 *arXiv:2501.13953*，2025d。

Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L. 等。Ro-borefer：面向机器人学视觉语言模型的空间指代与推理。*arXiv* 预印本 *arXiv:2506.04308*，2025。

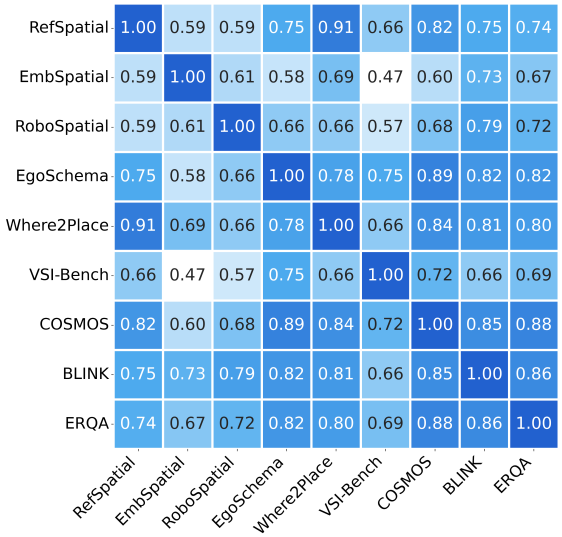


图4：基准间相似度热力图。结果显示，当前评估生态系统中存在严重冗余，基准测试之间的重叠率最高可达*91%*（例如，RefSpatial 与 Where2Place）。

A. 基准描述与冗余分析

本节提供有关源基准的补充信息，并展示评估格局中的冗余模式。

基准概览。表12 总结了本研究使用的所有基准测试，包括其描述和参考文献。

基准间冗余。图4 展示了各基准测试之间的两两语义相似度，揭示了当前评估生态系统中存在大量重叠。表8列举了不同基准测试间冗余样本的具体示例。

维度分布与覆盖范围。图5 展示了采用我们的抽样方法前后的能力分布。图6 展示了各维度样本嵌入的 UMAP 可视化结果，并用“+”标记我们选出的样本。

B. 详细评估设置

我们在下文提供不同模型系列的详细评估设置。所有模型均在 *A*800 上进行评估

Qwen3-VL模型。遵循 GitHub 官方建议，对于指令模型，我们为 Qwen3-VL 指令模型设置 top-p = 0.8、top-k = 20、温度 = 0.7 和重复惩罚 = 1.0。对于思考

模型	PercepObj	场景动作	空间几何	数量	可供性功能	PhysCaus	决策规划	动态场景	平均值
Qwen3-VL-235B-A22B-Thinking	62.23	60.03	71.23	65.22	68.58	65.68	68.26	66.56	65.97
InternVL-3.5-241B-A28B	61.45	62.61	70.99	60.23	62.30	67.56	71.58	68.68	65.68
GPT-5-20250807-Mini	64.80	65.23	72.17	61.90	73.22	60.58	63.20	63.07	65.52
Qwen3-VL-30B-A3B-Thinking	61.04	57.62	68.87	59.77	65.03	61.40	64.44	62.48	62.58
Internvl-3.5-38B	55.81	58.68	67.96	58.91	64.55	64.86	67.62	60.69	62.37
Qwen3-VL-235B-A22B-Instruct	64.44	62.69	54.95	57.90	66.39	62.72	65.08	60.90	61.88
Internvl-3.5-30B-A3B	54.08	58.67	69.81	57.84	54.10	66.36	68.10	65.37	61.79
Qwen3-VL-30B-A3B-Instruct	55.03	53.23	67.45	58.26	60.11	67.10	66.30	64.30	61.47
Internvl-3.5-8B	51.09	50.62	61.79	53.36	55.74	58.22	59.24	57.80	55.98
Qwen2.5-VL-72B-Instruct	57.48	56.03	49.06	51.93	67.21	45.60	55.16	55.61	54.76
Qwen2.5-VL-32B-Instruct	53.64	52.80	45.52	46.14	56.83	39.72	57.46	54.08	50.77
Qwen2.5-VL-7B-Instruct	49.05	51.02	43.16	47.42	52.46	40.52	50.84	51.30	48.22
Qwen2.5-VL-3B-Instruct	42.84	48.44	36.56	33.69	48.09	22.60	41.06	41.87	39.39

表7：受评模型在所有能力维度上的表现。所有分数均为百分比。“平均值”是各维度的平均值。

基准示例		
COSMOS	指示机器人“关闭抽屉”。 [...] 最有可能的下一步行动是什么？ (A) 向下方移动 [...]	
ERQA	机器人应该采取什么行动来关闭抽屉？ A. 向前移动 [...]	
RefSpatial-基准测试	指出粉色杯子后方、与红色碗距离相等的空位。应按以下格式 [...]	你的回答
Where2Place	找出蓝色马克杯和橙色碗之间的位置。你的答案应按 [...] 格式填写	
VSIBench	这个房间里有多少条毛巾？ 请回答问题 [...]	
ERQA	中有多少条不同的毛巾和/或垫子房间？ A. 两个 [...]	
VSIBench	哪个物体（椅子、背包、窗户、时钟）离门最近？ A. 椅子 [...]	
EmbSpatial-基准测试	哪个物体的距离最短？ A. 椅子 [...]	

表8：跨基准测试的四对冗余样本。

模型，我们使用：top-p $= 0.95$、top-k $= 20$、温度 $=0$。6、重复惩罚 $= 1.0$。最小像素数 $= 2$，097，152、最大像素数 $= 6$，291，456，以及总像素数 $= 128000 * 32 * 32 * 0.9$，适用于所有模型

Qwen2.5-VL 和 InternVL-3.5 模型。对于 Qwen2.5-VL 和 InternVL-3.5 系列，我们使用贪心解码，温度设为 0.01。请注意，InternVL-3.5-241B-A28B 使用 PyTorch 后端部署。

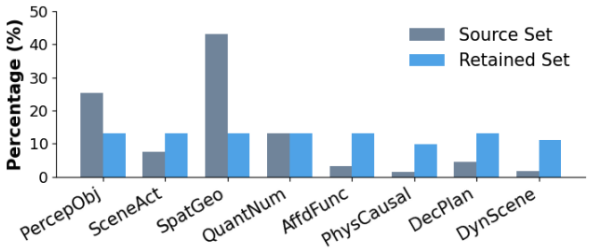


图5：多样性感知采样前后的跨维度能力分布。与高度偏斜的源分布相比，保留集合呈现出显著更加平衡的覆盖度。

C. 评估智能体的实现细节

C.1. 评估流程概述

本附录介绍了我们框架中使用的两个协作智能体的实现细节：(i) 数据智能体，它通过配备 5 个投票智能体的分配器构建能力维度和均衡基准；(ii) 评估智能体，它为模型评估合成可执行的推理和评分流程。我们将介绍完整的算法规范，以及不同智能体角色使用的提示词。

C.2. 数据智能体：维度构建与基准测试构建

C.2.1. 算法概述

数据智能体通过迭代式提议–审阅–分配和多样性感知采样，构建能力维度并建立结构化基准测试。完整流程见算法1。

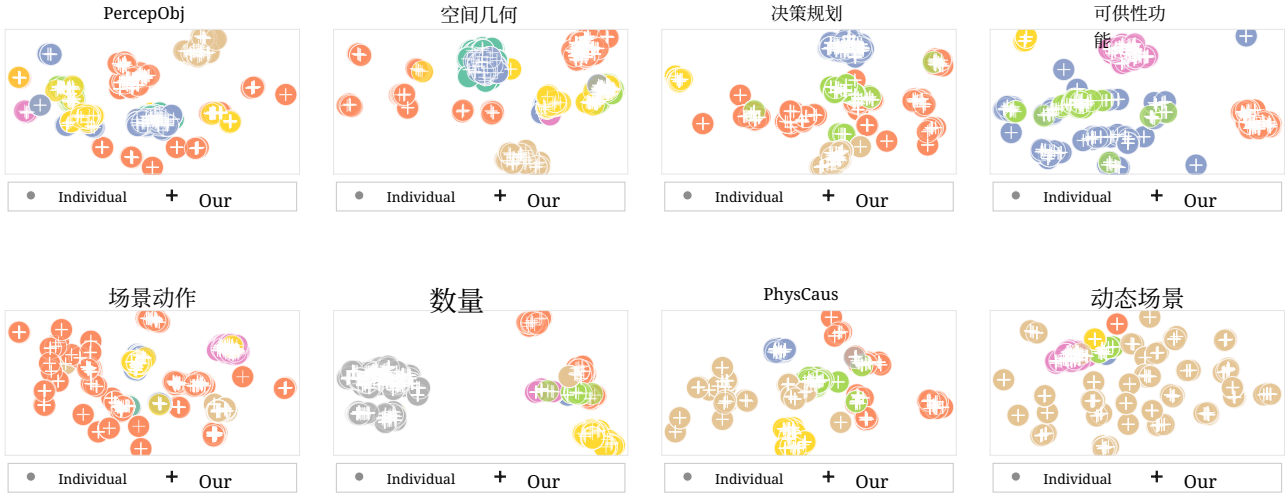


图6：所有维度中样本嵌入的 UMAP 可视化。

C.2.2. 数据智能体的角色

提议者 $\mathcal{A}_p$ 根据基准测试元数据 $\mathcal{B}$、审阅者提供的当前评审意见 C，以及记录先前提议和反馈的历史记忆 M，生成候选能力维度 D。借助 M，提议者可避免重复此前被否决的维度配置（例如冗余合并或不完整拆解），并依据C中的具体建议完善维度。

审阅者 $\mathcal{A}_r$ 使用分配器提供的平衡统计数据 S，评估候选维度 D 是否存在概念冗余、覆盖完整性不足及数据不平衡。提供结构化评议 C 以指导维度优化，例如建议合并冗余类别或补充缺失方面。当 S 显示存在严重不平衡（例如，某些维度的 d_k过小）$|E^{\mathrm{s}}_{d_k}|$ 时，审阅者会将这些信息纳入评议，以促使下一轮进行维度调整。

分配器 $\mathcal{A}_a$ **(与投票者** $\{\mathcal{A}_a^i\}_{i=1}^{N_v}$) 针对每个候选维度集 D执行两项关键操作：**(i) 维度分配** 通过由 N_v 名投票者组成的集成完成，他们各自使用来自 D 的维度为每个样本 $e_j \in \mathcal{E}$ 独立标注，最终分配结果由多数投票决定；**(ii) 多样性感知采样** 在每个维度 d_k 内使用基于嵌入的聚类对样本进行聚类，并从每个簇中选出代表样本，以最大限度地提高语义多样性，同时消除维度内冗余。计算平衡统计数据 $S = \{|E^{\mathrm{s}}_{d_k}|\}_{k=1}^{|D|}$，以识别代表不足的维度，并将其报告给审阅者，以便考虑合并或完善。

C.2.3. 数据智能体的提示设计

提议者的提示

现提供若干↪用于评估各种↪视觉语言↪模型（视觉语言模型）的基准测试描述。你的任务是综合这些来源↪并提出一套新的统一↪评估维度，专门↪ 针对具身智能体领域的视觉语言模型评估↪进行定制。

提供的信息

[基准测试描述开始]${description}[基准测试描述结束]

[历史记忆开始]{memory}[历史记忆结束]

[当前评审意见开始]{critique}[当前评审意见结束]

要求

1. 维度必须互不重叠、相互独立且合理。 2. 维度名称必须简洁，仅体现该维度的核心能力，并删除所有非必要的限定词。 3. 维度描述必须简洁、准确，并体现一种可泛化的能力。

4. 历史记忆部分（可选↪）记录以往的维度↪提案及其批评意见；请利用↪ 这些信息，避免↪重复使用被拒绝的配置↪或此前已发现的问题。

5. 当前评审意见部分为↪可选；如果存在，你必须↪结合评审反馈，↪通过调整、合并、细化↪或拆分维度，确保↪最终的维度集合能够解决任何↪被指出的弱点或重叠问题。↪

6. 输出必须采用整洁的 JSON 格式↪，不得附加任何评论。

{{ "name": "维度名称", "description": "用一到两↪句话说明该↪ 维度。"

}}

评审者提示词

你是一位专注且严谨的评估者↪，负责评估所提供的↪ 用于↪评估视觉语言模型↪（VLM）在特定领域中的维度框架。

所提供的信息

[领域描述开始]{domain}[领域描述结束]

[当前框架开始]{dimension_desc}[当前框架结束]

[平衡统计数据开始]{balance_stats}[平衡统计数据结束]

要求

1. 对以下内容进行文本评估 the ↪当前框架依据↪以下标准，并在提供↪平衡↪统计数据的情况下纳入相关反馈。a. 特异性：这些维度↪是否属于领域特定维度，而非↪一般模型属性（例如，是否↪ 不包括鲁棒性、↪效率、对未见过的↪物体或场景的处理能力、端到端↪能力或泛化能力）？

b. 最小化：该框架↪是否避免了不必要的碎片化？↪确定因强依赖关系、↪ 缺乏独立可评估性、↪或概念重叠而应当↪合并的维度。c. 平衡：如果提供了平衡统计数据↪（例如维度↪频率计数），请评估↪某些维度是否样本过少，↪以及是否应与相关维度合并，↪ 以确保↪足够的覆盖度。

2. 仅按以下格式返回一个 JSON对象，并为每项标准提供文本反馈：

{{ "特异性反馈": "", "最小化反馈": "", "平衡反馈": ""_

}}

投票者的提示词

你是某个特定领域的领域专家↪。你的任务是将一个↪新样本分配给最↪相关的预定义维度。

提供的信息

你将收到目标领域↪描述、当前的↪预定义评估维度↪以及样本本身。

[领域描述开始]{domain}[领域描述结束]

[预定义维度开始]{dimensions}[预定义维度结束]

[新示例开始]{example}[新示例结束]

要求

1. 根据维度名称和↪描述，确定最相关且最能↪ 体现样本所评估的主要能力的单一↪ 维度。2. 如果样本与任何↪维度都不太匹配，则将其归为“↪ 其他”。3. 按以下↪确切格式返回一个 JSON对象：

{{ "name": "维度_名称_或_其他", "reason": "解释该样本为何↪适配此维度" }}

C.3. 评估智能体：模型推理与评分逻辑合成

C.3.1. 算法概述

评估智能体通过沙箱执行器进行反馈引导的优化，合成可执行的推理逻辑和评分逻辑。完整流程如算法 2 所示。

C.3.2. 评估智能体的角色

评估器 $\mathcal{A}_e$ 合成推理逻辑 F^e，用于从 $\mathcal{T}$ 加载测试输入、在所需评估协议下执行目标模型 M，并生成预测结果 $\mathcal{P}$。根据沙箱执行器 $\mathcal{X}$ 提供的诊断反馈，迭代优化 F^e，直至执行可靠。

评分器 $\mathcal{A}_s$ 综合评分逻辑， F^s 根据测试用例 $\mathcal{T}$、模型预测$\mathcal{P}$和基准测试规范计算评估指标。它会根据来自 $\mathcal{X}$ 的反馈迭代优化 F^s，以确保指标一致且正确。

沙箱执行器 $\mathcal{X}$ 在隔离环境中执行生成代码（F^e 和F^s），并在执行失败时返回诊断反馈（例如运行时错误、API 不匹配），以便评估器和评分器进行迭代优化。

C.3.3. 评估智能体的提示设计

评估器的提示

你将编写一个视觉语言模型推理函数↪，该函数基于提供的视觉语言模型↪类型、基准示例和↪可选的附加信息。你的输出必须是完整、自包含的↪Python 函数，用于↪使用给定模型执行视觉语言模型推理，↪并使用给定模型路径、基准测试输入、↪ 模型生成参数和↪视觉参数。

提供给您的信息[模型类型开始]{vlm_类型}[模型类型结束]基准示例开始]
{基准_示例}[基准示例结束]

可选附加信息 此部分可能存在，也可能不存在。↪如果存在，请利用它来改进或↪修复你生成的代码。其包含：1. 之前生成的代码结果。2. 运行该代码时产生的执行错误消息，该代码运行于↪以下↪基准样例时。[之前生成的代码开始]{previous_code}[之前生成的代码结束]执行错误开始]{execution_error}[执行错误结束]

任务概述
你的目标是生成可靠且
↪功能完整的视觉语言模型推理
↪函数，该函数：
1. 仅加载指定的视觉语言模型模型
↪从给定的模型路径加载一次
↪然后再处理任何样本。
2. 接受由以下内容组成的基准测试输入列表：
↪字典，每个字典都包含一个问题
↪以及视觉输入（图像或视频），
↪ 并支持多样本处理
↪ 推理。
3. 处理通过 gen_kwargs 提供的模型生成参数↪（↪默认：max_new_tokens 1024，↪温度 0.01，随机种子 42）。
4. 应用来自以下内容的视觉参数
↪visual_kwargs（默认：max_frames↪ 32）。其中 max_frames 定义↪ 从视频中↪采样的最大帧数。如果视频↪包含的帧数少于 max frames_↪帧，则应使用全部帧；↪ 如果包含更多帧，则应↪ 均匀采样至↪max frames 帧。_
5. 对每个样本执行推理，并↪ 仅以字符串列表的形式↪返回模型的预测输出。
6. 使用以下函数签名：
def run_视觉语言模型_推理(模型_路径: str,
↪ 基准测试_输入：列表[字典]，
↪生成_参数：字典 = {{ "
↪ 最大_new_令牌": 1024 , "
↪温度": 0.01, "随机种子": 42 }},
↪ 视觉_kwargs: dict = {{ "
↪max_帧": 32 }}) -> list[str]
7. 如果提供了可选的补充信息
↪请使用这些信息来优化、 调试，
↪ 或纠正你的代码生成。
8. 代码不得使用任何 try/except
↪块或其他异常处理
↪ 结构。

9. 必须在推理过程中显示多样本↪进度条（例如 tqdm）。10. 强制要求以下输出格式：↪ 仅生成一个包含实现代码的 Python↪ 代码块，↪代码块外不得包含任何解释、注释↪或↪ 文本。

评分器的提示词

你将根据提供的信息编写基准测试评分↪函数，其中包括↪基准信息、基准↪示例以及可选的附加↪信息。你必须输出一个完整且自包含的↪Python 函数，用于↪依据该基准测试的↪评分方法计算分数。

提供给你的信息[基准信息开始]
{benchmark_infor}[基准信息结束][基准示例开始]{benchmark_example}[基准示例结束]

可选附加信息 本部分可能出现，也可能不出现。↪如果出现，请利用它来改进或↪修正生成的评分逻辑。其中包含：1. 先前生成的评分↪ 函数。2. 在基准样例上运行该函数时产生的↪执行错误信息。↪[先前生成的代码开始]{previous_code}[先前生成的代码结束][执行错误开始]{execution_error}[执行错误结束]

任务概述
你的目标是生成一个可靠且↪功能完备的评分函数↪ ，该函数应：1. 根据↪基准测试名称和示例↪推断出正确的评分规则。2. 正确处理不同的评估↪ 类型（例如，精确匹配↪题、数值题、↪多项选择题，或任何↪特定于基准测试的格式）。3. 针对单个预测答案对↪ 计算并返回一个浮点分数。

4. 严格使用以下函数签名↪ ： def eval_score(pred, answer) -> int 5. 确保函数自包含，↪不依赖外部文件，也不使用↪任何 try/except 代码块或其他↪异常处理结构。6. 如果提供了↪可选的附加信息，请用它来完善、调试↪或修正生成的评分↪ 函数。7. 强制采用以下输出格式：↪ 只输出一个包含实现代码的 Python 代码块↪，不得在代码块↪ 外添加任何说明、注释↪或文字↪。

D. 人类—智能体一致性协议

本附录详述用于验证分配器的自动维度分配与人类专家理解相一致的人在回路验证协议。

D.1. 标注设计

我们并未要求标注员从零开始分配维度——那需要他们完全熟悉整个分类体系并带来巨大的认知负担——而是采用将强模型预标注与专家人工验证相结合的两阶段标注流程。

阶段1：模型预标注。 强参考模型（GPT-4o）根据完整的分类体系定义为每个样本进行初步维度标注，提出最相关的能力维度。

第二阶段：人工验证。人工标注员审查这些模型提出的标签，在必要时纠正错误，并生成经过最终验证的维度分配。这种设计在保留严格人工监督的同时，显著减轻了标注员的认知负担，使标注员能够专注于判断细化和边界案例，而非分类体系回忆。

D.2. 标注员招募

人工标注由 5 位领域专家完成，包括具备具身智能相关背景的博士研究人员和工程师。所有标注员在标注前都会接受统一的校准流程，并使用一致的维度定义和代表性样本，以确保对能力边界的理解一致。标注过程中，模型预测仅作为参考；标注员独立作出判断，并在存在歧义或明显分类错误时修订模型输出

维度	Cohen 的 κ	IAA
PercepObj	0.81	0.83
场景动作	0.82	0.84
空间几何	0.79	0.80
数量	0.77	0.78
动态场景	0.75	0.76
可供性功能	0.80	0.82
物理因果	0.76	0.77
决策规划	0.78	0.79
平均值	**0.78**	**0.80**

表 9：采用人在回路验证协议对能力维度分配进行的人机一致性。Cohen's κ 衡量自动分配结果与人工验证标签之间超出偶然水平的一致性，而 IAA 则报告五位专家标注员之间的标注者间一致性。

观察到这种情况。

D.3. 标注流程

对于每个维度，我们构建一个平衡的评估集，其中包含由数据智能体提出的候选正样本和候选负样本。所有样本均经过完全随机化，并以盲测方式呈现：标注员不会获知数据智能体的预测结果、基准测试来源或类别平衡情况。每个样本由多名标注员独立审核，最终的维度标签通过多数投票确定。对于每个维度，我们抽取 100 个样本进行评估。

D.4. 一致性指标

我们衡量三种一致性：

• **Cohen 的** κ：自动分配结果与人工验证标签之间超出偶然水平的一致性，并考虑了随机一致性。

• **标注者间一致性（IAA）**：五位专家标注员之间的 Fleiss κ，用于衡量人类判断的一致性。

D.5. 结果与分析

表9 总结了人机一致性结果。在所有维度上，自动维度分配与经人类验证的标签之间均表现出高度一致性，平均 Cohen's κ 为 0.78，平均标注者间一致性为 0.80，表明一致性达到较高水平，接近完美。这些结果表明，自动推导出的能力维度与专家的人类判断高度吻合。

E. 以人为中心的评估协议

本附录详细介绍以人为中心的评估流程。我们的目标是构建具身规划器的人类感知排名，并将其与新基准所产生的排名进行比较，以检验基准测试信号是否符合人们对真实具身场景中模型能力的直觉。

E.1. 目标

我们构建人类感知评估信号，作为自动化基准测试的互补轴：标注员观看真实世界视频，并评判模型具身行动计划的质量。随后，我们将所得的以人为中心的模型排名与新基准排名进行比较，以验证（i）新基准是否反映人类感知的能力差异，以及（ii）所归纳的能力维度是否能以更低的冗余更好地区分模型优势。

E.2. 场景

我们使用八种真实世界视频场景：**家庭**、**杂货店**、**学校**、**办公室**、**仓库**、**医院**、**工作坊**和**实验室**。每段视频时长约为 15–30 秒，每个场景包含 4–6 个以自然语言指令描述的子任务。

E.3. 一个具体示例：家庭场景

我们提供一个具体的家庭示例来说明评估设置。家庭指令如下：

> "我要去上班了。请把鞋子放到鞋架上，把桌子上的垃圾扔进垃圾桶，并把沙发上的脏衣服放进洗衣机。"

这条指令被拆解为三个子任务：**(1)鞋子整理**：识别鞋子和鞋架，导航至每只鞋子旁，抓取鞋子并将其放到合适的鞋架位置，同时合理摆放成对鞋子并保持朝向；**(2)桌面垃圾处理**：识别桌子并区分垃圾和非垃圾物品，导航至桌边，逐一抓取垃圾物品，导航至垃圾桶并将其投入，直到看不到剩余垃圾；**(3)沙发衣物装载**：识别沙发区域和脏衣服，导航并抓取衣服，找到洗衣机及其打开方向，导航至洗衣机，打开机门，将衣服放入滚筒并关上机门，重复此过程，直到沙发上的衣服全部清理完毕。

该示例以一段简短的视频片段（15–30 秒）和指令的形式呈现给模型，并要求模型生成一份明确的分步计划，在物理上

可行且与可见内容一致。

E.4. 模型输入与所需输出格式 at

所有模型接收相同的输入，包括：(i) 场景视频，(ii) 全局系统提示，以及 (iii) 包含任务指令的用户提示。我们为所有模型采用统一的输出结构，以确保结果具有可比性：

系统提示（全局）。

你是一名具身人工智能规划器。你将收到一段展示当前环境的短视频。你的任务是理解场景、识别相关物体，并生成完整且物理上可行的行动计划，以实现目标。你必须仅依据视频中可见或可推断的信息。不得假定视频中未展示的物体或位置存在。不得跳过必要的物理步骤（例如打开容器、靠近、伸手）。严格遵循所需的输出结构。

用户提示

[提供的环境视频]
请仅根据视频中可见的内容，生成一份明确、逐步展开的具身行动计划。
你的回答必须明确说明以下方面的推理：(1) 场景元素和物体位置，(2) 任务目标和子目标，(3) 行动顺序和时间依赖关系，(4) 空间关系和导航轨迹，(5) 物体状态及其随时间发生的状态变化，(6) 适用时相关物体的数量和计数。

请按以下结构组织你的输出：
1. 场景理解 2. 目标解读 3. 行动计划（分步） 4. 最终计划（精简格式）

E.5. 人工评分维度与权重

标注员根据视频和模型的回答，为每个模型输出打分。我们使用**八项标准**，每项按**1–5李克特量表**评分，并以**等权重（各12.5%）**汇总：

- **规划**：计划是否涵盖关键步骤并完成指令；
- **场景理解**：模型是否正确感知并描述场景结构；
- **动态一致性**：计划是否随时间推移仍与物体状态/位置保持一致；
- **物体接地**：所引用的目标物体是否与视频中的内容相符（例如，是否为正确的物品）；
- **物理推理**：动作是否在物理上可行并尊重物体状态；
- **定量推理**：计数/数量相关的决策是否正确（例如物品数量）；
- **空间推理**：空间关系以及导航/轨迹是否被正确使用；
- **时间推理**：行动顺序是否合理，并遵循先决约束。

E.6. 人工评估流程

对于每个场景，标注员依次：(1) 观看视频，(2) 阅读指令，(3) 阅读经过匿名化处理的模型输出，以及 (4) 按照八个维度为每项输出打分。每个模型与场景的组合至少由 3–5 名标注员独立评分。场景和模型的顺序均经过随机化，以减少顺序偏差。表 13以一个典型的家庭场景为例，具体展示了人工评估标准及相应的模型输出，并逐项列出各维度的评估结果。

E.7. 分数汇总与排名对齐

对于每项模型输出，我们通过对八个维度分数取平均值来计算加权总分（每个维度的分数均归一化到相同尺度，并赋予相同权重）。随后，我们对各标注员的评分以及所有场景的评分分别取平均值，从而得到每个模型最终的以人为中心的分数，并据此得出人类感知排名。

与此同时，我们还计算同一组模型在新基准测试下的基准分数和排名。我们使用 Spearman's ρ 和 Kendall'sτ 量化基准测试所得排名与以人为中心的排名之间的一致性。高相关性表明，新基准测试捕捉到的评估信号与人类判断一致，也支撑了这样一种观点：所推导出的能力维度反映了人类感知到的具身能力。

F. 详细保真度分析

本节详细分析第 4.6 节中评估智能体的保真度评估。我们分别分析推理保真度和评分保真度。

F.1. 推理保真度

表 10 报告了由评估智能体生成的推理逻辑保真度。 F^f 我们使用智能体生成的输出与参考模型结果之间的归一化编辑距离来衡量保真度：保真度$_{\text{inf}} = 1 -$编辑距离$_{\text{norm}}$，其中编辑

维度	Q2.5-7B	Q3-8B	IVL3.5-8B	平均值
PercepObj	92.5	92.4	96.5	93.8
场景动作	92.0	91.3	92.5	91.9
空间几何	94.5	93.4	94.7	94.2
数量	94.9	93.8	94.4	94.4
可供性功能	91.8	91.1	92.3	92.7
PhysCaus	91.6	91.9	92.4	92.0
决策规划	96.6	95.6	96.9	96.4
动态场景	93.9	93.4	93.7	93.6
平均值	93.5	92.9	94.2	**93.6**

表10：智能体生成的推理输出与参考模型结果之间的推理保真度。

距离通过除以可能的最大距离（较长序列的长度）归一化为 $[0,1]$。该智能体在所有能力维度和模型配置上的平均推理保真度为 93.6%。

F.2. 评分保真度

表 11 展示了评分逻辑 F^s 由评估智能体生成的保真度。我们使用相对分数差衡量保真度：保真度$_{\text{score}} = 1 - \frac{|\text{分数}_{\text{pred}} - \text{分数}_{\text{ref}}|}{\text{分数}}$，其中分数$_{\text{pred}}$ 和分数$_{\text{ref}}$ 分别表示由智能体生成的评分函数与参考评分函数计算得到的分数。

所有能力维度和模型配置的平均评分保真度均超过 97%。少数差异源于接近数值阈值的边界情况，而非生成的评分逻辑存在根本性错误。

维度	Q2.5-7B	Q3-8B	IVL3.5-8B	平均值
PercepObj	94.9	94.5	93.5	94.3
场景动作	98.4	98.5	99.0	98.6
空间几何	95.8	97.9	96.5	96.7
数量	95.1	94.0	95.7	94.9
可供性功能	99.6	97.0	100.0	98.9
PhysCaus	100.0	99.1	100.0	99.7
决策规划	100.0	98.3	100.0	99.4
动态场景	100.0	96.4	100.0	98.8
平均值	97.9	97.0	97.9	**97.9**

表11：智能体生成的评分函数与参考评分函数之间的评分保真度。

算法 1 数据智能体

输入： 基准信息 $\mathcal{B} = \{B_i\}_{i=1}^{N_B}$；基准测试样本 $\mathcal{E} = \bigcup_{i=1}^{N_B} \widetilde{E_i}$ **输出：** 基准测试 $\mathcal{B}$ 及其维度 D；分配到各维度的样本 $\mathcal{E}^{\text{s}}$ **角色：** 提议者 $\mathcal{A}_p$；审阅者 $\mathcal{A}_r$；分配器 $\mathcal{A}_a$ （投票者$\{\mathcal{A}_a^1, \ldots, \mathcal{A}_a^{N_v}\}$） **初始化：** 记忆 $M \leftarrow [\,]$、批评意见 $C_0 \leftarrow$ ”“、平衡统计数据 $S^{(0)} \leftarrow \emptyset$

重复

(a) 维度归纳 **重复**$D_{i+1} \leftarrow \mathcal{A}_p(\mathcal{B}, C_i, M)$ $C_{i+1} \leftarrow \mathcal{A}_r(D_{i+1}, S^{(r)})M \leftarrow M \cup \{(D_{i+1}, C_{i+1})\}$ $i \leftarrow i+1$**直到** $D_{i+1} = D_i D^{(r+1)} \leftarrow D_i$**(b.1) 维度分配** **对每个示例** $e_j \in \mathcal{E}$ **执行** **对每个投票者** $\mathcal{A}_a^k$, $k = 1, \ldots, N_v$ **执行**$D_{jk} \leftarrow \mathcal{A}_a^k(e_j, D^{(r+1)})$ **结束循环** $D_j \leftarrow$**多数投票(**$\{D_{jk}\}_{k=1}^{N_v}$**) 结束循环** **对每个维度** $d_k \in D^{(r+1)}$ **执行**$E_{d_k} \leftarrow \{e_j \mid d_k \in D_j\}$**结束循环** **(b.2) 多样性感知采样** **对每个维度** $d_k \in D^{(r+1)}$ **执行**$X_{d_k} \leftarrow \{$**编码器(**e**)：将** $e \in E_{d_k}\}$**聚类为** K **组，质心为** $\{\mu_j\}_{j=1}^{K} E^{\text{s}}_{d_k} \leftarrow \{$ **arg min**$_{x \in C_j} \|x - \mu_j\|$ **：** $j = 1, \ldots, K\}$**结束循环**$S^{(r+1)} \leftarrow \{|E^{\text{s}}_{d_k}|$ **：** $d_k \in D^{(r+1)}\}r \leftarrow r+1$**直到** $D^{(r+1)} = D^{(r)}$ $D \leftarrow D^{(r)}\mathcal{E}^{\text{s}} \leftarrow \bigcup_{k}^{|D|}$

$_{=1} E^{\text{s}}_{d_k}$

维度	描述
VSI-Bench（Yang等人，2025a）	VSI-Bench包含三类共八项任务：空间构型、测量估计和时空。空间构型任务（物体计数、相对距离、相对方向、路线规划）测试模型对空间布局的理解，这对人类来说更加直观。
EgoSchema（Mangalam 等人，2023）	EgoSchema 是一项诊断基准，用于评估现代多模态系统对超长视频语言的理解能力。理解时长较长的自然视频需要多种相互关联的能力，例如行动与场景理解、感知并追踪物体状态、长期视觉记忆、抽象推理、层次化信息聚合，以及more.
COSMOS（Azzolini 等人，2025）	COSMOS 关注：（1）“任务完成验证”：判断任务或子任务是否已成功完成的能力；（2）“行动可供性”：评估执行特定行动或进行某项操作是否可以朝目标推进；(3)“下一个合理动作预测”：识别最合理的下一步行动或子任务，以朝着指定目标推进的能力。
Where2Place（Yuan 等人，2025）	Where2Place 使用空间关系评估自由空间指代。这些图像收集自各种杂乱环境。每张图像都标注了一句描述所需空闲空间以及所需区域掩码的句子区域。
ERQA（Team 等，2025b）	ERQA 是一项基准测试，专门关注具身智能体与物理世界交互时可能需要的能力。ERQA 包含400 道视觉问答（VQA）风格的选择题，涵盖涵盖多种类别，包括空间推理、轨迹推理、动作推理、状态估计、指向、多视角推理和任务推理。
EmbSpatialBench（Du 等人，2024年）	EmbSpatialBench，一个用于评估具身环境中的大型视觉语言模型空间理解能力的基准测试。重点关注六种空间关系从自我中心视角出发，包括上方、下方、左侧、右侧、近处和远处，全面涵盖坐标的三个维度。
OmniSpatial（Jia 等人，2025）	OmniSpatial 在其四个主要类别中提供了具有代表性的样本的空间推理。视角转换：任务展示了从不同视角理解空间关系的能力。动态推理：这包括涉及理解物体运动及其变化的任务。空间交互：这些任务侧重于与空间环境进行互动复杂逻辑：此类别涵盖更高层次的推理，例如模式识别——（风格、数量、属性、位置）以及几何推理（多面体展开、截面与投影、心理旋转、组装、分析几何——学）。
RefSpatialBench(Zhou 等人，2025)	RefSpatialBench 包含 200 张真实世界图像，并经过人工标注用于物体位置和放置的任务。填补了评估空间指代与多步推理方面的空白。
RoboSpatial (Song 等人，2025)	RoboSpatial：一种用于评估视觉的全新空间推理基准测试——用于机器人学真实室内环境的视觉语言模型。它由——350 个空间推理问题及众包收集的 RGBD图像组成这些图像由配备深度传感器的手持 iPhone 相机拍摄。每张图像都标注了三类空间关系问题——空间配置、空间上下文和空间兼容性，从而全面评估——机器人应用中的空间理解。空间配置：确定物体的相对位置（例如，“杯子是否位于笔记本电脑的左侧？”）。空间上下文：识别相对于参照物的空闲区域（例如，“识别碗左侧的空闲空间。”）。空间兼容性：评估物体是否能够适配指定区域（例如，“椅子能否放在书桌前面？”）。
BLINK（Fu 等人，2024年）	BLINK评估了更全面的视觉感知能力，如多视角推理、深度估计和反射率估计。

表 12：基准测试及其描述。

维度	评估标准（家庭场景）	示例模型输出（qwen3-vl-235b）
规划	该计划完全完成了所有子任务的规定要求-在人类指令中指定。具体来说，计划是否包括：（1）将所有 鞋子从地板上放到鞋架上，（2）处理所有 垃圾物品（例如纸杯和牛奶盒），将它们从桌子上放进垃圾桶，以及（3）将所有 脏衣服从沙发上转移走放入洗衣机，然后明确关上洗衣机盖?	“靠近茶几 → 收集零食包装 → 丢进垃圾桶 → 收集空的杯子 → 把杯子摆放在桌子上 → 移到鞋子区域 → 捡起地上的鞋 → 放到鞋子架子 → 验证整洁状态……靠近咖啡桌——从沙发区出发，向前移动约 *1.5* 米即可到达桌子……朝白色柜子旁的鞋子区域走去，移动约 *3* 米穿过房间……”
场景理解	计划是否正确感知并提及环境中可见的所有关键元素，包括左侧的鞋架、沙发前的桌子、垃圾桶、放着脏衣服的沙发，以及右侧的洗衣机，且没有臆造未见物体或假设存在无依据功能?	“视频展示了一间现代客厅，里面有木地板和几个主要物件：- 一张白色*L*形沙发，曾有一位穿红色衣服的人坐在上面 - 一张茶几，上面摆放着多件物品：零食包装（薯片）、两个白色物品：零食包装（薯片）、两个白色杯子，一本书和一个容器 - 一个黑色铁丝垃圾桶摆放在茶几旁边 - 一个黑色金属鞋架，上面有*3-4*双鞋（一些放在架子上，一些放在地板上） - 一台洗衣机靠近灰色窗帘 …”
动态一致性	计划是否保持一致的物体状态以及整个执行过程中的位置？尤其是，一旦鞋子被放置到架子上，或衣服被移入洗衣机，这些物体是否不再被指称为仍留在其原始位置的后续quent步骤?	“拿起空的零食包装，从茶几表面（可见*2-3*个包装）……移动零食包装至位于*0.5* 米处茶几左侧的垃圾桶……将零食包装放入垃圾桶……”
物体定位	该计划是否能将行动正确定位到场景中预期的目标物体，例如区分桌子上的垃圾物品与非垃圾物体，并识别地板上的鞋子与沙发上的衣服 ，没有 con-将它们与附近但无关的物品融合在一起?	“捡起茶几表面上的空零食包装（可见 *2-3* 个）……移动零食包装，放进位于茶几左侧 *0.5* 米处的垃圾桶……”
物理推理	计划是否遵守物理约束，并遵循物体的可供性，例如在装入衣服之前打开洗衣机盖，使用合适的身体部位进行抓握，并避免做出违背物理规律或不安全的行为吗?	“把杯子放回茶几上，以一种有条理的方式（因为没有水槽可见于场景）…… 走向鞋子区域附近的白色柜子，大约移动*3*米横穿房间……”
定量推理	计划是否正确推理了数量在场景中，例如需要捡起多少只鞋子有多少件垃圾物品（例如纸——杯子或牛奶盒）在桌子上，以及沙发上所有可见的衣服是否都被转移，而不是将多个物体视为一个整体行动?	“拿起咖啡桌上的两个空白色杯子……拿起地板上的鞋子，它们在鞋架旁边（可见 *3-4* 双）……”
空间与轨迹	该计划是否正确推理了空间关系与导航轨迹，包括鞋子是否被放置到鞋架上（而不仅仅是靠近它），以及是否衣服被放置进洗衣机，沿着合理的移动路径，在位置之间移动?	“把零食包装移到位于茶几左侧 *0.5* 米处的垃圾桶里——……从沙发区接近茶几……朝白色鞋子区域走去柜子……”
时间推理	计划是否遵循时间依赖关系例如先导航再进行动作先抓握，再打开容器以供使用，以及完成一个子任务后再转而执行接下来进行下一步，且不交错执行未完成的步骤?	“一个黑色金属鞋架，上面有 *3-4* 双鞋子（有些在架子上，有些在地板上）……捡起鞋架旁地板上的鞋子鞋架……将每双鞋子放到黑色金属鞋架上合适的架子上……”

表13：家庭场景的分维度人工评估示例。

算法 2评估智能体

输入： 基准信息 $\mathcal{B}$；测试样例 $\mathcal{T} \subseteq \mathcal{E}^{\mathrm{s}}$；模型 M；能力维度 D；基准测试样本 $\mathcal{E}^{\mathrm{s}} = \bigcup_{k=1}^{N_d} E^{\mathrm{s}}{}_{d_k}$
输出： 推理逻辑 F^e；评分逻辑 F^s；模型预测结果 $\{\mathcal{P}_{d_k}\}_{d_k \in D}$；模型得分 $\{S_{d_k}\}_{d_k \in D}$
角色： 评估器 $\mathcal{A}_e$；评分器 $\mathcal{A}_s$
沙盒： 执行器 $\mathcal{X}$

阶段 I：模型推理逻辑构建
$F_0^e \leftarrow \mathcal{A}_e(M, \mathcal{T})$
重复
  $E_i \leftarrow \mathcal{X}(F_i^e$<风格 id='3'>,</风格> M<风格 id='6'>,</风格> $\mathcal{T})$
  $F_{i+1}^e \leftarrow \mathcal{A}_e(M, \mathcal{T}, F_i^e, E_i)$
  $i \leftarrow i + 1$
直到 $E_i =$ 无（无错误反馈）
$F^e \leftarrow F_i^e$

第二阶段：模型评分逻辑构建
$\mathcal{P}_{\mathcal{T}} = F^e(M, \mathcal{T})$
$F_0^s \leftarrow \mathcal{A}_s(\mathcal{T}, \mathcal{P}_{\mathcal{T}}, \mathcal{B})$
重复
  $E_j \leftarrow \mathcal{X}(F_j^s, \mathcal{T}, \mathcal{P}_{\mathcal{T}})$
  $F_{j+1}^s \leftarrow \mathcal{A}_s(\mathcal{T}, \mathcal{P}_{\mathcal{T}}, \mathcal{B}, F_j^s, E_j)$
  $j \leftarrow j + 1$
直到 $E_j =$ 无（无错误反馈）
$F^s \leftarrow F_j^s$

第三阶段：基准评估执行
对 $d_k \in D$ **执行**
  $\mathcal{P}_{d_k} \leftarrow \mathcal{X}(F^e, M, E^{\mathrm{s}}_{d_k})$
  $S_{d_k} \leftarrow \mathcal{X}(F^s, E^{\mathrm{s}}_{d_k}, \mathcal{P}_{d_k})$
结束 for