AirCopBench: A Benchmark for Multi-drone Collaborative Embodied Perception and Reasoning

Jirong Zha [1*], **Yuxuan Fan**[2*], **Tianyu Zhang**[3], **Geng Chen**[4], **Yingfeng Chen**[5], **Chen Gao**[6 †], **Xinlei Chen**[1 †]

[1]Shenzhen International Graduate School, Tsinghua University
[2]The Hong Kong University of Science and Technology (Guang Zhou)
[3]School of Electrical and Electronic Engineering, Nanyang Technological University
[4]College of Software, Jilin University
[5]Weiyang College, Tsinghua University
[6]BNRist, Tsinghua University
zhajirong23@mails.tsinghua.edu.cn, yfan546@connect.hkust-gz.edu.cn, tianyu016@e.ntu.edu.sg, chengeng0201@gmail.com, chenying24@mails.tsinghua.edu.cn, chgao96@gmail.com, chen.xinlei@sz.tsinghua.edu.cn
[*] Equal contribution. [†] Corresponding authors.

主讲人：郭周鹏
日期：12.11

1 Introduction

背景：多模态大模型在单智能体视觉任务表现优异，但在多无人机协同感知领域缺乏评估基准；多机系统具备覆盖广、鲁棒性强等优势，亟需填补这一空白 。

动机：现有基准多聚焦于高质量图像的基础感知，难以评估模型在现实感知退化（如噪声、遮挡）条件下的复杂具身协同能力，故需构建更具挑战性的航空基准 。

现有方法局限：

- **感知局限：** MLLMs在单一任务（OD、Seg)表现出色，但在**协作**感知任务中存在研究空白。
- **感知任务设置简单：** 缺少多样的**退化场景**数据（如，遮挡，传感器噪声、能见度差、数据丢失和环境干扰）
- **缺乏具身推理：** 缺少自我中心的**具身协作**感知**推理**

假设： 通过构建包含多种现实退化场景与分层级语义任务的基准，系统揭示 MLLMs 在多视图协同感知中的短板。

贡献：

- **提出新颖任务体系：** 设计了包含场景理解、物体理解、感知评估及协同决策 4 大维度、14 项具体任务的评估框架 。
- **构建挑战性数据集：** 创建了首个涵盖遮挡、噪声等多种感知退化场景的航空协同数据集（2.9k+ 多视图图像），包含事件级与物体级标注 。
- **生成大规模 VQA：** 结合模型、规则与人工方法，生成并清洗了 14.6k+ 高质量视觉问答对，用于全面测试模型能力 。
- **广泛评估与验证：** 评测了 40 个主流 MLLMs，揭示了其协同感知能力显著落后于人类，并通过微调实验证实了模拟数据的有效性 。

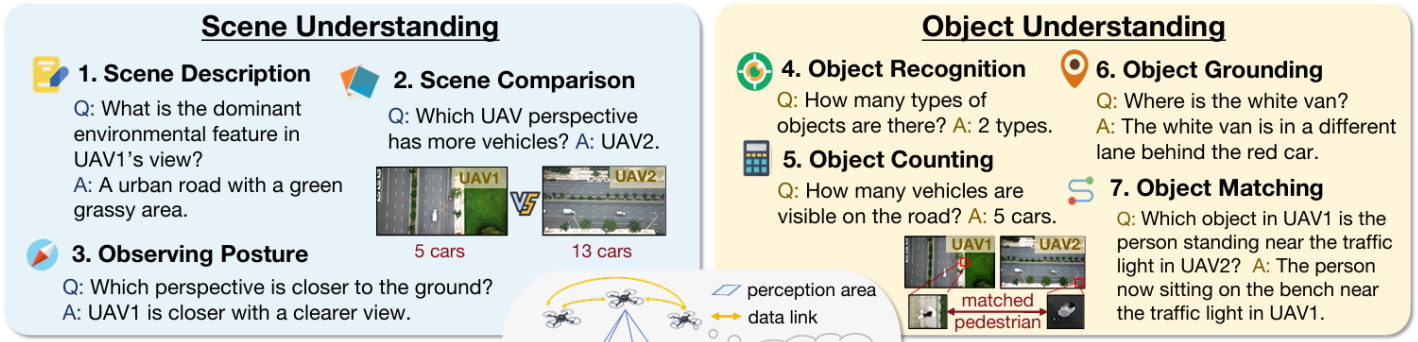
2 Benchmark Tasks

场景理解： 弄清楚“我看到了什么环境”以及“我是怎么看的”。（场景描述、比较、姿态）

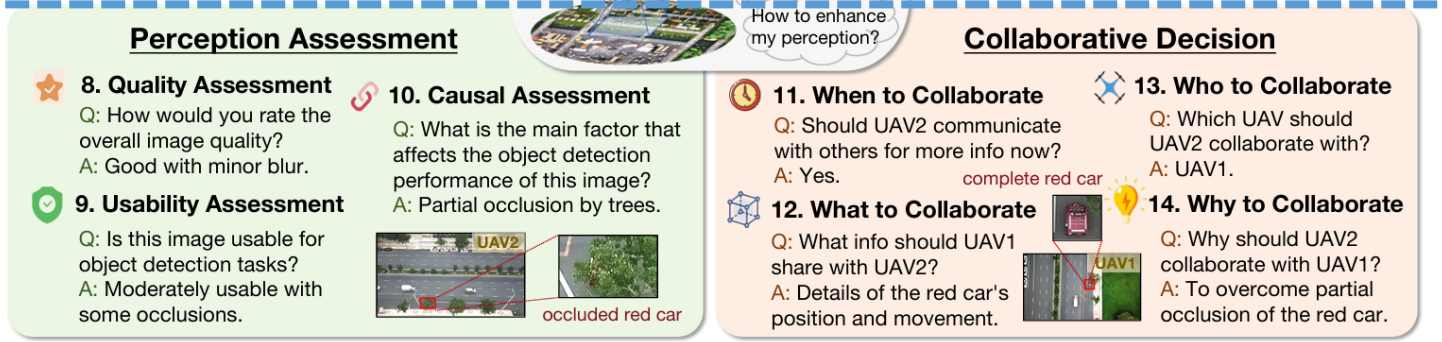
例子： “无人机1视野里的主要环境特征是什么？” → “是一条有绿化带的城市道路。”
例子： “哪个视角的车更多？” → “无人机2。”
例子： “哪个视角离地面更近？” → “无人机1更近且视野更清晰。”

物体理解： 聚焦具体目标，解决“是什么、有多少、在哪儿”以及“能不能对上号”的问题。（物体识别、定位、计数、匹配）

例子： “这里有几种物体？” → “2种。”
例子： “白色面包车在哪？” → “在红车后面的另一条车道上。”
例子： “无人机2红绿灯旁的人，对应无人机1里的哪个？” → “无人机1里坐在长椅上的那个人。”



例子： “你给这图质量打几分？” → “良好，只有轻微模糊。”
例子： “这图能用来做物体检测吗？” → “勉强能用，有一些遮挡。”
例子： “影响这张图检测效果的主要因素是什么？” → “被树木部分遮挡了。”



例子： “无人机2现在需要和其他人通信获取信息吗？” → “需要。”
例子： “无人机1应该分享什么信息给无人机2？” → “红车的具体位置和移动细节。”
例子： “无人机2应该找哪架无人机合作？” → “无人机1。”
例子： “为什么无人机2要找无人机1？” → “为了解决红车被部分遮挡的问题。”

感知评估： 自我诊断，判断“我的眼睛（传感器）好不好使”以及“看不清是为什么”。（质量可用性、因果评估）

协同决策： 团队配合，决定“要不要摇人”、“找谁帮忙”以及“交换什么情报”。（何时协同、协同什么，协同对象，协同原因）

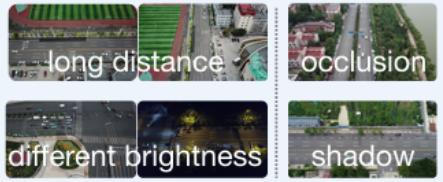
2 Benchmark Generation

Data Collection

Simulator data
3- / 5- / 6-UAV group

multi-view collaboration

Real-world data

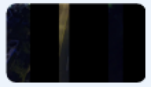


challenging degradations

Derived data

- Noise injection
  - sensor failure

- Partial masking
  - data loss



Data Annotation

Event-level labeling

- Image quality

Very poor | Poor | Fair | Good | Excellent

- Perception usability

Yes | No

- Perception degradation

occlusion | shadow | ...

- Collaborative analysis

when | what | who | why

Object-level labeling

- Object list <...,...,...,...>
- Bounding box $<x_1,y_1,w_1,h_1,...>$
- Target attribute <...,...,...>

Question Generation

Model-based generation

- Divided task
- Role-playing
- CoT prompt
- Few-shot

Q: Why should UAV4 collaborate with another UAV?
A: To overcome building occlusion and gain a more complete view of the scene.

Rule-based generation

Q: Which UAV perspective shows more vehicles?
A: UAV2.

{"anno1": "8 objects (car: 5, bicycle: 1, person: 2)",
"anno2": "19 objects (car: 13, person: 4, bicycle: 2)"}

Human-based generation

Q: Which UAV perspective is closest to the drone target? D.Equally close.

A. B. C.

Quality Control

Standard examination

Scoring criteria:

- Required content ⭐
- Format consistency ⭐⭐
- Answer validity ⭐⭐⭐
- Question length ⭐⭐⭐⭐✓

Blind filtering

VQAs Models $A_1 = Y_1 = \cdots = Y_n$?

$[Q_1, A_1]$ M_1 Y_1
$[Q_2, A_2]$ M_2 Y_2
$[Q_m, A_m]$ M_n Y_n

Save | no | yes | Delete

Human refinement

- Ambiguous questions
- Invalid options
- Incorrect answers

Data Collection

Data Annotation

Question Generation

Quality Control

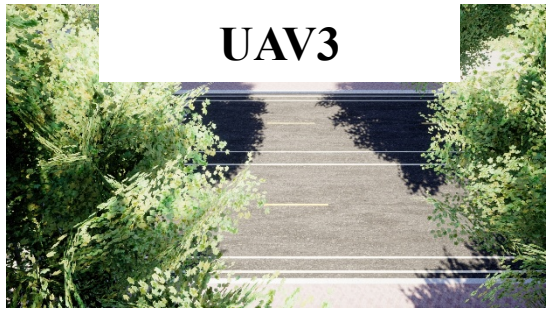
2 Benchmark Generation

Real

sim

noise

loss

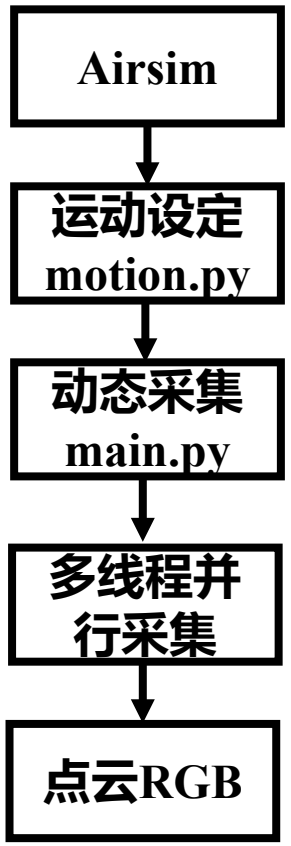


UAV1

UAV2

UAV3

虚拟数据采集流程



Data Collection

Data Annotation

Question Generation

Quality Control

2 Benchmark Generation

enhancing precise identification in complex scenes.

Event-level Labeling. To emphasize the role of collaboration in perception, we introduce novel "event" annotations to assist in generating multi-UAV interaction strategies, such as when and why inter-UAV communication is needed, who is suitable for information retrieval, and what observation information should be shared. Despite labeling for collaborative decision analysis, the annotations in this part also include image quality scoring, perception usability assessment, and perception degradation reasoning for better event understanding. The whole manual annotation process costs over 200 hours. More details are in *Appendix*.

Object-level Labeling. Traditional annotations for object labeling involve the list of specified objects in the scene, the 2D/3D bounding box of each object, and the corresponding attributes of objects like motion state (Hu et al. 2022). This enables accurate collaborative perception for targets, including target detection, classification, and tracking.

事件级标注：
直接记录多机当前感知的事实状态，与专家的判断逻辑（Groundtruth）。

Quality: "Excellent"（对应质量评估）
Collaboration_when: "1 (Yes)"（对应何时协同）
Collaboration_who: "UAV2"（对应协同对象）
Degradation: "Small target"（对应因果评估）

base

对象级标注：
检测框

Data Collection

Data Annotation

Question Generation

Quality Control

2 Benchmark Generation


```
{
  "img1": "23-00000001-UAV1.jpg",
  "id": 1,
  "Quality": "Excellent (5/5) - Sharp, clean, balanced colors, no artifacts.",
  "Usibility": "1 (Available)",
  "Object_type": {
    "choices": [
      "Vehicles",
      "Pedestrians"
    ]
  },
  "PerceptionIssues": [
    {
      "x": 60.31496062992125,
      "y": 75.777194517352,
      "width": 2.7296587926509304,
      "height": 2.7996500437445206,
      "rotation": 0,
      "rectanglelabels": [
        "Too small"
      ],
      "original_width": 1920,
      "original_height": 1080
    }
  ],
  "Collaboration_when": "1 (Yes)",
  "Collaboration_why": {
    "choices": [
      "The current image suffers from object occlusion, while the other perspective has a complete view of the object and can provide complementary information.",
      "The target appears too small or distant in the current image, but is larger and more visible in the other persepctive.",
      "Current image misses the target due to out-of-view, but other views may capture it."
    ]
  },
  "Object_count": "7",
  "Collaboration_who": "UAV2",
  "Degradation": "Small target / Long distance",
  "annotator": 1,
  "annotation_id": 1,
  "created_at": "2025-06-24T06:17:08.956105Z",
  "updated_at": "2025-06-29T17:36:58.178063Z",
  "lead_time": 375.567
},
```


8帧一标注

场景内容： Object_type（目标类别）, Object_count（目标数量）。
图像评估： Quality（清晰度评分）, Usibility（任务可用性判断）。
缺陷定位： Degradation（退化类型，如“小目标”），PerceptionIssues（缺陷区域的精确坐标 x,y,w,h 及标签）。

协同决策： When（时机）、Who（对象）、What（动机）

Data Collection

Data Annotation

Question Generation

Quality Control

2 Benchmark Generation


```
{
  "sequence_frame": "23-00000001",
  "question_id": "MDMT_when2col_UAV1_1",
  "question_type": "4.1 When to Collaborate (UAV1)",
  "question": "Should UAV1 collaborate with another UAV to address need for collaboration due to incomplete information?",
  "options": {
    "A": "Yes, due to partial occlusion of key objects",
    "B": "No, the scene is fully visible",
    "C": "Yes, due to poor visibility of the objects",
    "D": "No, all objects are clearly captured"
  },
  "correct_answer": "A",
  "uav_paths": {
    "UAV1": "All_Samples/UAV1/23-00000001-UAV1.jpg",
    "UAV2": "All_Samples/UAV2/23-00000001-UAV2.jpg"
  }
},
```


"MDMT_when2col_UAV1_1"
"MDMT_what2col_UAV1_1"
"MDMT_who2col_UAV2_2“
"MDMT_why2col_UAV2_7“
"MDMT_CA_UAV1_173“….

问题生成：
Method1：基于规则
Method2：基于大模型
Method3：人工标注

Real数据只有两个UAV
Sim有三、四个

Data Collection

Data Annotation

Question Generation

Quality Control

2 Benchmark Generation

问题生成：
Method1：基于规则
Method2：基于大模型（使用GPT-4o：def call_chatgpt_api(messages, retries=3):）
Method3：人工标注（在Data Annotation部分实现）

任务 4.1: 何时协同 (When to Collaborate)
 # [基于规则生成] # 优先使用人工标注的 'Collaboration_when' 字段。
 # 逻辑：读取 JSON 中的 "Yes/No"，直接构造对应的选择题。
result_when = CALL generate_rule_based_collaboration_when_q(annotation)
IF result_when IS EMPTY **(无标注数据):**
 # [基于大模型生成]
 # 兜底方案（Fallback）。
 # 逻辑：将图像输入 GPT-4o，让模型自行判断是否需要协同。
result_when = CALL generate_model_based_collaboration_when_q(images)
ADD result_when TO uav_questions

Data Collection

Data Annotation

Question Generation

Quality Control

2 Benchmark Generation

质量把控：
Method1：标准把控
Method2：难度把控
Method3：人工把控

Method2.难度把控（Blind Filtering）：
使用 n 个 MLLM 在不输入图像的情况下预测答案，如果都能答对则删除该问题。并未给出代码。

Method3.人工把控：
人工修改模糊问题、无效选项或错误答案

1.内容完整性：evaluate_question_quality:(检查必要字段是否缺失)

```
required_fields = ["question_id", "question_type", "question", "options", "correct_answer"]
for field in required_fields:
    if field not in result:
        issues.append(f"Missing required field: {field}")
```

2.格式一致性：evaluate_question_quality:(检查必要字段是否缺失)

```
if "options" in result:
    # ... checks if options is a dictionary
    # ... checks if len(options) == 4
    # ... checks if keys are ["A", "B", "C", "D"]
```

3.答案有效性：evaluate_question_quality:(检查必要字段是否缺失)

```
#检查正确答案是否在选项键
if correct not in options:
    issues.append("Correct answer must be one of the option keys...")
# 计算文本相似度，防止生成雷同选项
similarity = difflib.SequenceMatcher(None, opt1, opt2).ratio()
if similarity > 0.85: return False...
```

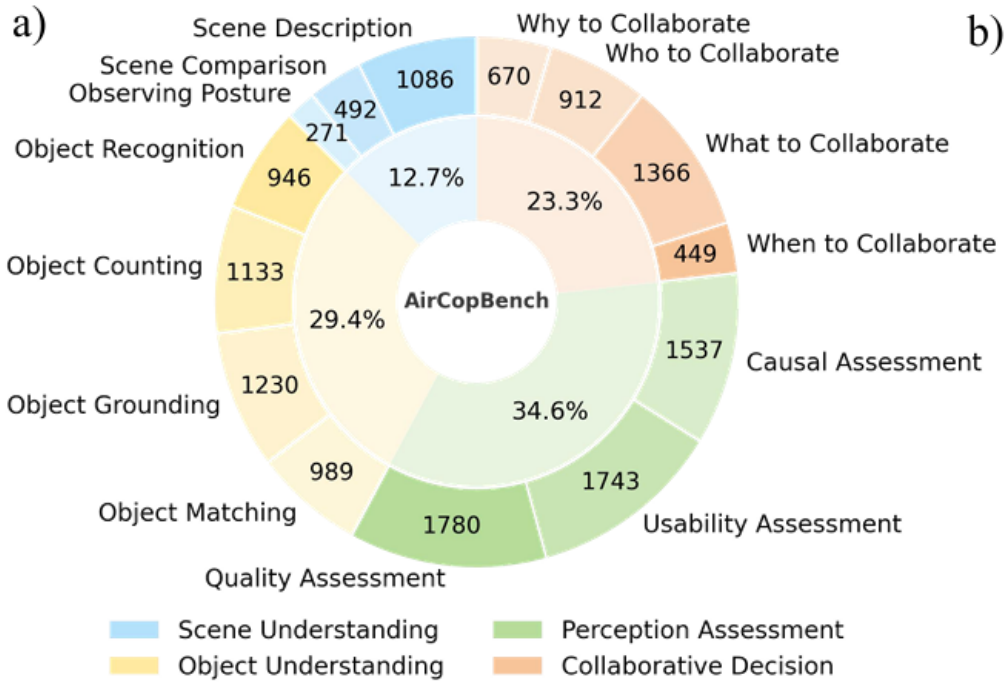
Data Collection

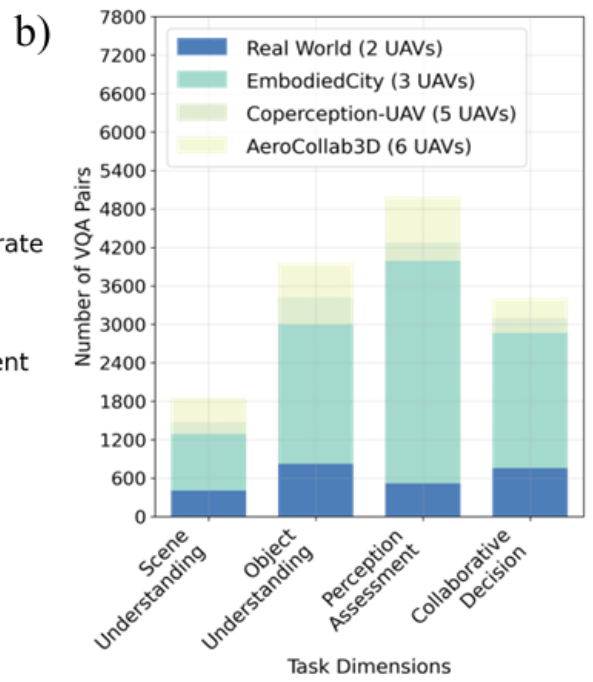
Data Annotation

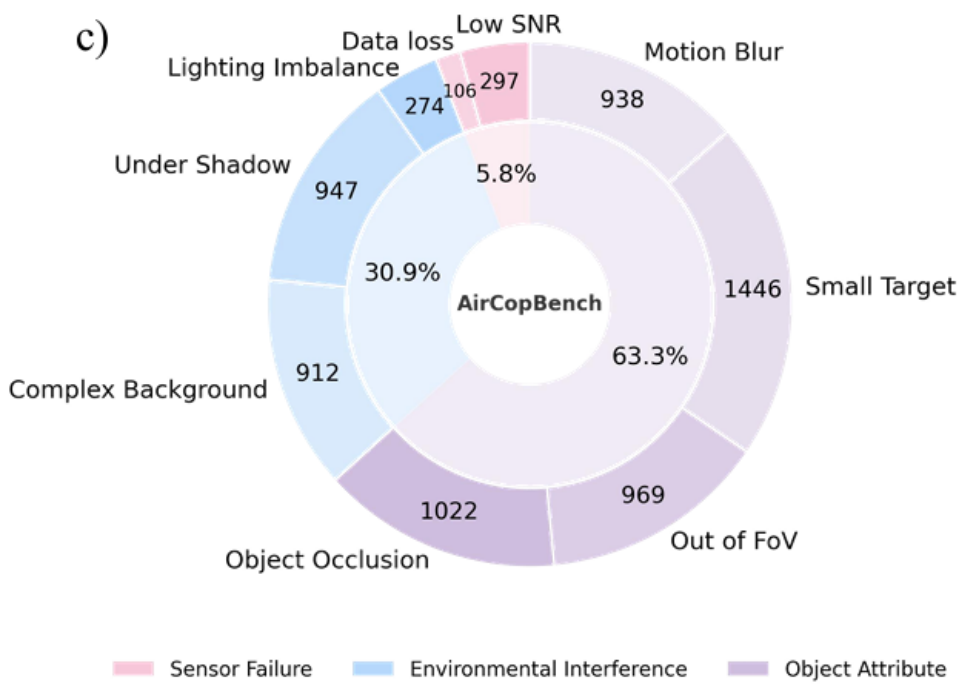
Question Generation

Quality Control

2 Benchmark Statistics







任务数据分布概览

具体数据量对比

退化数据分布

VQA pairs

3 Experiments

1.评价指标：准确率
答对题目数/总题目数
按：任务类型、真实虚拟数据分别打分。

2.对比设定：
Proprietary Models（闭源）
Open-source Models（开源）
微调

3.训练： LLaMA-Factory
Easy and Efficient LLM Fine-Tuning

4.测试：
VLMEvalKit/evaluation.py

Method	Rank	Avg.	Scene Understanding			Object Undstanding				Perception Assessment			Collaborative Decision			
			Scene Desc.	*Scene Comp.*	*Obs. Post.*	*Obj. Rec.*	*Obj. Cnt.*	*Obj. Grnd.*	*Obj. Mtch.*	*Qual. Ass.*	*Usab. Ass.*	*Caus. Ass.*	*When Coll.*	*What Coll.*	*Who Coll.*	*Why Coll.*
Baseline																
Random	-	23.47	19.30	44.19	18.52	16.67	23.46	27.68	17.14	19.51	19.51	28.57	41.38	18.52	24.69	24.69
Human	-	78.25	71.43	75.86	42.86	85.71	88.89	83.04	87.62	90.48	91.46	85.71	51.72	82.72	82.72	75.31
Proprietary Models (API)																
GPT-4o-2024-11-20	2	51.79	64.91	55.81	**44.44**	65.48	**44.44**	50.89	**67.62**	29.76	48.78	70.24	34.48	58.02	14.81	60.49
Gemini-2.5-Pro	5	49.08	70.18	62.79	37.04	**67.86**	27.16	47.32	47.62	15.48	36.59	**72.62**	41.38	61.73	34.57	**65.43**
Claude-Sonnet-4-20250514	3	50.73	59.65	55.81	33.33	61.90	33.33	52.68	56.19	35.71	**57.32**	71.43	20.69	60.49	30.86	51.85
Qwen-Max-VL-latest	4	50.53	52.63	65.12	29.63	61.90	41.98	54.46	61.90	**44.05**	46.34	66.67	17.24	53.09	**39.51**	39.51
Step-1o-turbo	1	**52.87**	**75.00**	**70.83**	21.05	66.10	33.33	**61.54**	59.26	27.42	55.93	71.67	41.38	**67.27**	18.52	56.60
Doubao-seed-1-6-flash-250615	2	51.79	59.65	48.84	37.04	54.76	**44.44**	53.57	63.81	41.67	52.44	67.86	**48.28**	54.32	34.57	48.10
Open-source Models																
Phi-4-multimodal-instruct	5	52.76	63.16	60.47	33.33	51.19	25.93	52.68	65.71	26.19	40.24	70.24	24.14	**66.67**	70.37	60.40
Qwen2.5-VL-7B-Instruct	10	47.33	66.67	60.47	25.93	63.10	25.93	50.89	51.43	47.56	47.56	66.67	13.79	34.57	25.93	43.21
Qwen2.5-VL-72B-Instruct	4	54.90	59.65	65.12	33.33	58.33	**41.98**	63.39	**67.62**	48.78	48.78	73.81	17.24	59.26	37.04	48.15
InternVL3-8B	6	52.18	56.14	60.47	25.93	59.52	30.86	58.04	56.19	56.10	51.22	71.43	20.69	54.32	53.09	50.62
InternVL3-78B	3	55.38	66.67	**67.44**	44.44	64.29	24.69	**67.86**	62.86	**58.54**	**58.37**	**76.19**	13.79	55.56	50.62	41.98
Janus-Pro-7B	12	44.91	52.63	48.84	22.22	51.19	18.52	58.04	51.43	28.57	46.34	61.90	31.03	60.49	33.33	37.00
Chameleon-7B	15	38.22	36.84	37.21	44.44	25.00	24.69	29.46	46.67	16.67	20.73	53.57	27.59	45.68	**75.31**	49.30
PaliGemma-3B	17	24.25	19.30	37.21	22.22	30.95	35.80	18.75	13.33	11.90	21.95	47.62	**65.52**	17.28	16.05	16.05
MiniCPM-V2.6	7	51.99	63.16	62.79	33.33	**65.48**	40.74	49.11	49.52	46.43	48.78	66.67	41.38	58.02	46.91	45.68
Ovis2-16B	1	**59.17**	**68.42**	**67.44**	29.63	64.29	28.40	56.25	**67.62**	58.33	57.32	66.67	51.72	60.49	60.49	**71.60**
Ovis-U1-3B	16	37.34	57.89	46.51	22.22	41.67	29.63	36.61	39.05	27.38	45.12	63.10	24.14	24.69	29.63	25.93
Kimi-VL-A3B-Thinking	2	56.84	59.65	60.47	25.93	61.90	38.27	58.04	63.81	45.24	50.00	**76.19**	48.28	**66.67**	51.85	62.96
Mimo-VL-7B-RL	9	48.59	61.40	58.14	29.63	64.29	34.57	53.57	57.14	46.43	53.66	75.00	10.34	50.62	17.28	33.33
LLaVA-NeXT-7B-hf	14	38.31	28.07	46.51	18.52	35.71	25.93	39.29	52.38	29.76	37.80	59.52	27.59	55.56	25.93	29.63
LLaVA-NeXT-13B-hf	13	39.28	31.58	44.19	33.33	42.86	30.86	40.18	46.67	40.48	41.46	50.00	34.48	44.44	34.57	24.69
Skywork-R1V3	8	48.94	46.15	43.33	**46.67**	41.51	40.00	50.00	46.99	40.74	56.60	67.27	33.33	52.94	48.08	51.92
mPLUG-OWL3	11	47.14	57.89	60.47	22.22	50.00	25.93	56.25	41.90	27.38	47.56	55.95	44.83	54.32	50.62	54.32
XComposer-VL-7B	18	23.26	14.81	20.00	18.75	25.45	26.42	18.82	22.62	27.78	13.79	13.79	12.50	24.07	39.62	33.96
Fine-tuned Models																
LLaVA-NeXT-13B	3	57.61	40.35	60.47	25.93	52.38	45.68	59.82	60.95	57.14	62.20	69.05	37.93	58.02	70.37	66.67
Qwen-2.5-VL-7B	1	**74.30**	63.16	**65.12**	**33.33**	**69.05**	**75.31**	**66.07**	**72.38**	**76.19**	**82.93**	**83.33**	**55.17**	**77.78**	**91.36**	**85.10**
Qwen-2.5-VL-3B	2	66.44	**73.68**	55.81	**33.33**	59.52	34.57	57.14	62.86	66.67	73.17	82.14	**55.17**	**77.78**	90.12	80.20
Sim-to-Real Experiments																
Qwen2.5-VL-7B	-	47.77	50.00	55.56	11.11	83.33	50.00	27.78	55.56	61.11	44.44	82.35	10.53	38.89	64.71	27.70
AirCop-7B	-	67.41	50.00	77.78	11.11	83.33	88.89	77.78	77.78	77.78	94.44	82.35	31.58	50.00	76.47	50.00

3 Experiments

1. 模型综合表现 (Model Comparison)整体挑战巨大:

- AirCopBench 对所有 MLLMs 均构成显著挑战，即便是表现最好的 Ovis-16B 模型，准确率也仅为 59.17% 。这揭示了当前模型在具身感知和协同决策能力上的不足 。
- 开源模型崛起：顶级开源模型（如 Ovis2-16B, Kimi-VL-A3B）的表现已匹配甚至略微超越顶尖的
- 闭源专有模型 。但两者在**多图像推理**上均存在**短板** 。
- 任务能力偏科：模型在基础的“场景描述”上表现尚可，但在需要领域知识和目标导向推理的任务（如“可用性评估”、“何时协同”）上表现不佳，亟需提升适应性决策能力 。

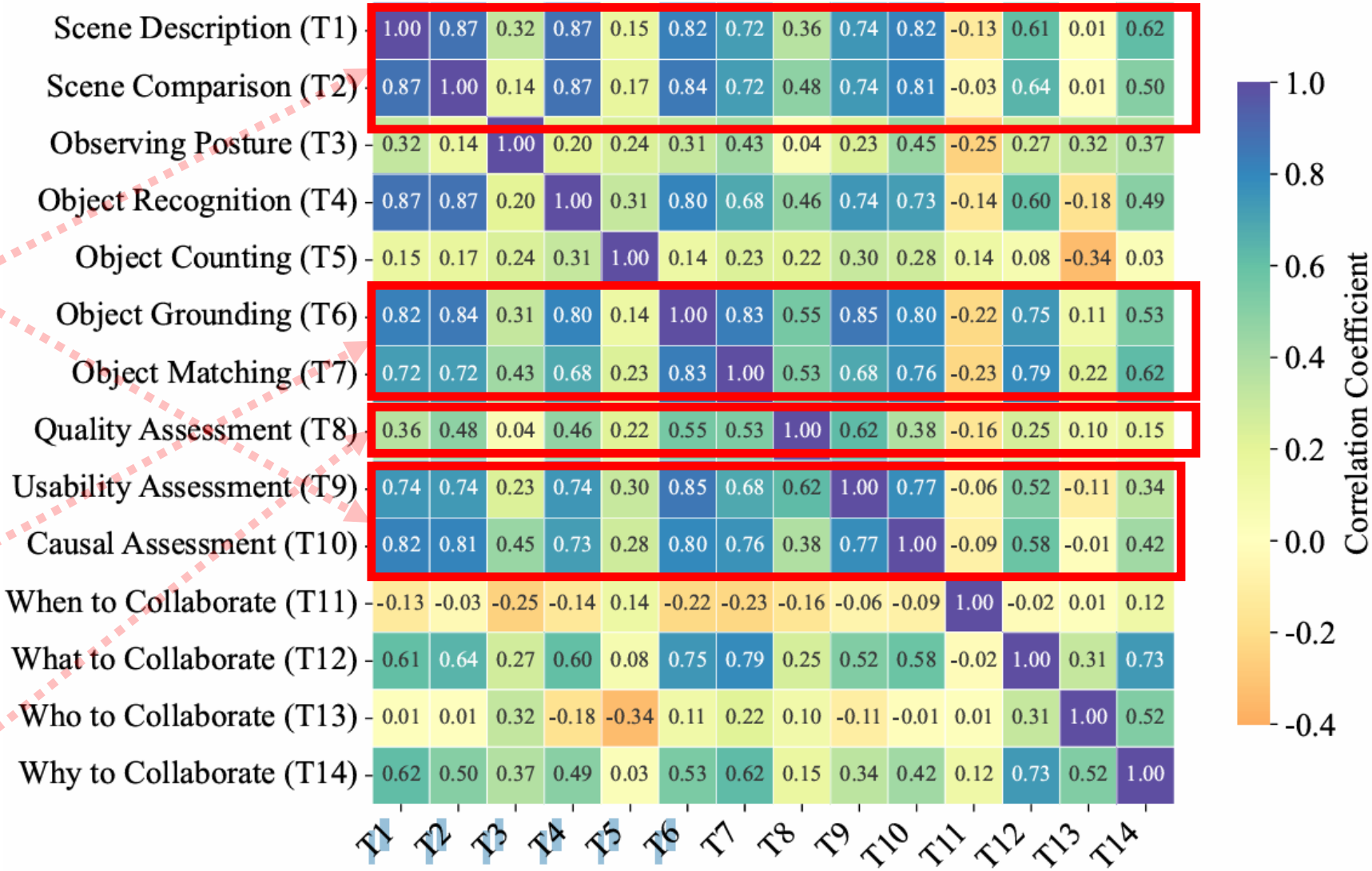
Method	Rank	Avg.
Baseline		
Random	-	23.47
Human	-	78.25
Proprietary Models (API)		
GPT-4o-2024-11-20	2	51.79
Gemini-2.5-Pro	5	49.08
Claude-Sonnet-4-20250514	3	50.73
Qwen-Max-VL-latest	4	50.53
Step-1o-turbo	1	**52.87**
Doubao-seed-1-6-flash-250615	2	51.79
Open-source Models		
Phi-4-multimodal-instruct	5	52.76
Qwen2.5-VL-7B-Instruct	10	47.33
Qwen2.5-VL-72B-Instruct	4	54.90
InternVL3-8B	6	52.18
InternVL3-78B	3	55.38
Janus-Pro-7B	12	44.91
Chameleon-7B	15	38.22
PaliGemma-3B	17	24.25
MiniCPM-V2.6	7	51.99
Ovis2-16B	1	**59.17**
Ovis-U1-3B	16	37.34
Kimi-VL-A3B-Thinking	2	56.84
Mimo-VL-7B-RL	9	48.59
LLaVA-NeXT-7B-hf	14	38.31
LLaVA-NeXT-13B-hf	13	39.28
Skywork-R1V3	8	48.94
mPLUG-OWL3	11	47.14
XComposer-VL-7B	18	23.26
Fine-tuned Models		
LLaVA-NeXT-13B	3	57.61
Qwen-2.5-VL-7B	1	**74.30**
Qwen-2.5-VL-3B	2	66.44
Sim-to-Real Experiments		
Qwen2.5-VL-7B	-	47.77
AirCop-7B	-	67.41

3 Experiments

2. 任务相关性分析 (Correlation Analysis)

- 因果评估是核心：“因果评估”任务与几乎所有其他任务均呈高相关性，表明理解感知退化的原因（Causality）是具身认知的关键因素 。
- 单视能力是基础：“物体匹配”与“物体识别/定位”高度相关，说明有效的多视角理解依赖于强大的单视角感知能力 。
- 质量评估是复合任务：“质量评估”与其他任务中度相关，说明它需要综合场景理解、物体识别和匹配能力来进行决策 。

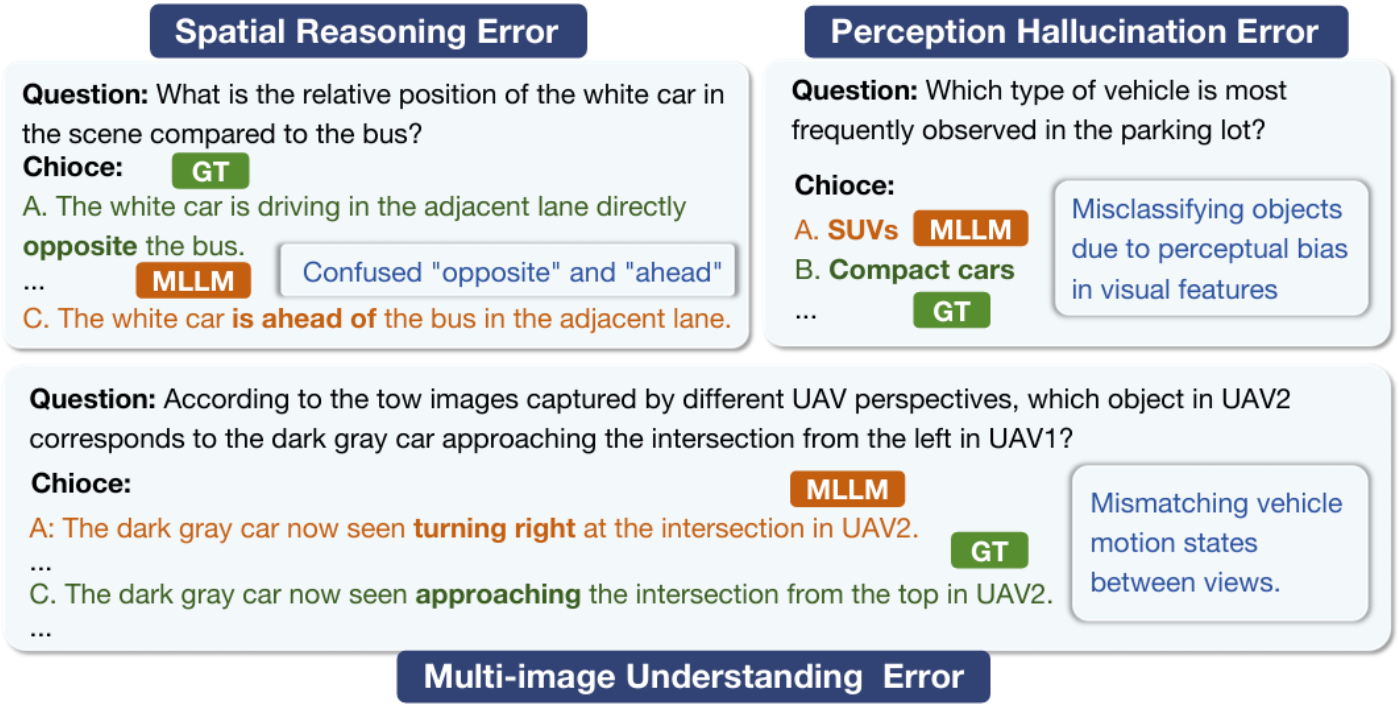
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



3 Experiments

3. 微调与迁移有效性 (SFT & Sim-to-Real)

- **监督微调 (SFT) 提升显著：**在 AirCopBench 上进行微调后，Qwen-2.5-VL-7B 的准确率提升了 +26.97%（达到 74.30%），证明了数据集能有效增强模型的协同感知能力 。
- **虚实迁移 (Sim-to-Real) 可行：**使用模拟器数据训练的模型（AirCop-7B）在真实世界数据上的表现提升了 +19.64%（从 47.77% 升至 67.41%），验证了模拟数据的有效性和良好的泛化能力 。



4. 错误类型归因 (Error Analysis)

- **感知幻觉 (Perception Hallucination)：**错误识别或虚构物体。
- **空间推理错误 (Spatial Reasoning Error)：**搞错物体的位置关系或方向。
- **多图理解错误 (Multi-image Understanding Error)：**在跨视角比较、匹配或整合信息时出现逻辑错误。

Thanks!