The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

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

Abstract

Multimodal Large Language Models (MLLMs) have shown promise in single-agent vision tasks, yet benchmarks for evaluating multi-agent collaborative perception remain scarce. This gap is critical, as multi-drone systems provide enhanced coverage, robustness, and collaboration compared to single-sensor setups. Existing multi-image benchmarks mainly target basic perception tasks using high-quality single-agent images, thus failing to evaluate MLLMs in more complex, egocentric collaborative scenarios, especially under real-world degraded perception conditions.To address these challenges, we introduce AirCopBench, the first comprehensive benchmark designed to evaluate MLLMs in embodied aerial collaborative perception under challenging perceptual conditions. AirCopBench includes 14.6k+ questions derived from both simulator and real-world data, spanning four key task dimensions: Scene Understanding, Object Understanding, Perception Assessment, and Collaborative Decision, across 14 task types. We construct the benchmark using data from challenging degraded-perception scenarios with annotated collaborative events, generating large-scale questions through model-, rule-, and human-based methods under rigorous quality control. Evaluations on 40 MLLMs show significant performance gaps in collaborative perception tasks, with the best model trailing humans by 24.38% on average and exhibiting inconsistent results across tasks. Fine-tuning experiments further confirm the feasibility of sim-to-real transfer in aerial collaborative perception and reasoning.

Project — https://embodiedcity.github.io/AirCopBench/

1 Introduction

Multimodal Large Language Models (MLLMs) have transformed artificial intelligence (AI) by enabling powerful processing of diverse inputs, such as text and images (Zhang et al. 2025b). While MLLMs have excelled in single-agent vision tasks like object detection and semantic segmentation (Saini et al. 2025), as well as general multi-image understanding tasks (Cheng et al. 2025; Bai et al. 2025b), their performance in collaborative perception with multiple unmanned aerial vehicles (UAVs) remains underexplored (Wang et al. 2025d). In multi-UAV systems, drones work together to tackle complex, dynamic visual tasks through information exchange, offering enhanced coverage, robustness, and flexibility compared to single-agent configurations (Wang et al. 2022; Zha et al. 2024).

Despite the potential of multi-UAV systems, benchmarks evaluating MLLMs in collaborative perception are scarce. Current evaluations focus primarily on single-agent vision tasks, failing to address the unique challenges of multi-UAV collaborative perception, such as obstacle occlusion (Kil et al. 2024; Xiao et al. 2025). Furthermore, existing benchmarks on collaborative perception (Wang et al. 2024c; Tian et al. 2024) in Tab. 1 face two major limitations that hinder their applicability in adaptive real-world scenarios:

- *Oversimplified Perception Setups*: Despite occlusion, numerous factors such as sensor noise, poor visibility, data loss, and environmental interference can degrade the quality of UAVs' observation data, thus reducing overall perception accuracy (Wang et al. 2025b). Current benchmarks consider only a limited set of degradation types, restricting their adaptability to real-world conditions.
- *Lack of Embodied Reasoning*: Aerial collaborative perception demands rich semantic information processing and highly adaptive cooperation in complex environments. Traditional non-egocentric, programmatic decision-making schemes, based on global data fusion, hinder UAVs from making context-aware, human-like, first-view decisions, limiting their ability to understand their state, adapt to changes, and collaborate efficiently (Zhao et al. 2025b).

Therefore, it is essential to assess MLLMs' cognitive abilities in embodied perception and collaborative decision-making under challenging perceptual conditions, as illustrated in Fig. 1. More related work can be found in *Appendix*.

---

[*]Equal Contribution.

[†]Project Leader.

[‡]Corresponding Author.

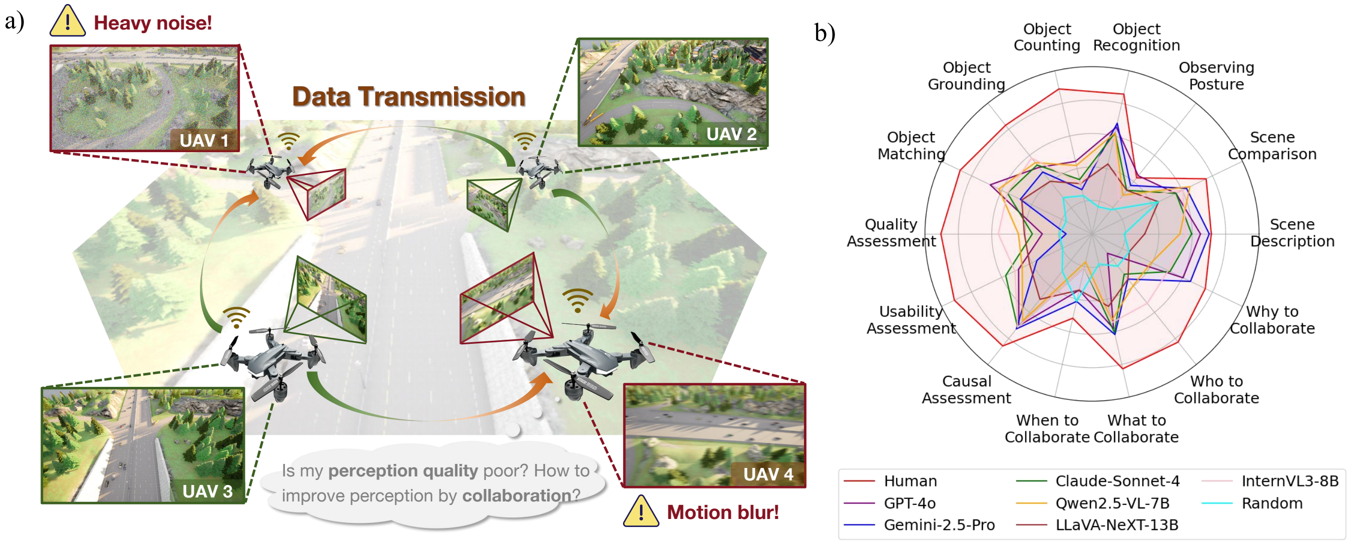


Figure 1: a) Illustration of multi-drone collaborative perception with various perception degradation. b) The performance of 6 popular MLLMs, along with human and random guess baselines, on AirCopBench.

Benchmark	Modality	Object Category	Perception Degradation	VQA	VQA Num.	Embodied
Coperception-UAV (Hu et al. 2022)	RGB	Vehicles	/	✗	/	✓
MDMT (Liu et al. 2023)	RGB	Vehicles, bicycles, pedestrians	Occlusion, lighting, blur	✗	/	✓
UAV3D (Sunderraman, Ji et al. 2024)	RGB	Vehicles	/	✗	/	✓
AeroCollab3D (Tian et al. 2024)	RGB	Vehicles, pedestrians	/	✗	/	✓
Air-Co-Pred (Wang et al. 2024c)	RGB	Vehicles, bicycles, pedestrians	Occlusion, long distance	✗	/	✓
MUIRBENCH (Wang et al. 2024a)	RGB, text	Vehicles, buildings	Occlusion	✓	2.6k	✗
All-Angles Bench (Yeh et al. 2025)	RGB, text	Pedestrians	Occlusion	✓	2.1k	✗
UrBench (Zhou et al. 2025)	RGB, text	Vehicles, buildings	/	✓	11.6k	✗
Our AirCopBench	RGB, text, point cloud	Drones, vehicles, bicycles, pedestrians	Occlusion, shadow, noise, lighting, out of FoV, data loss, long distance/small target, blur	✓	14.6k	✓

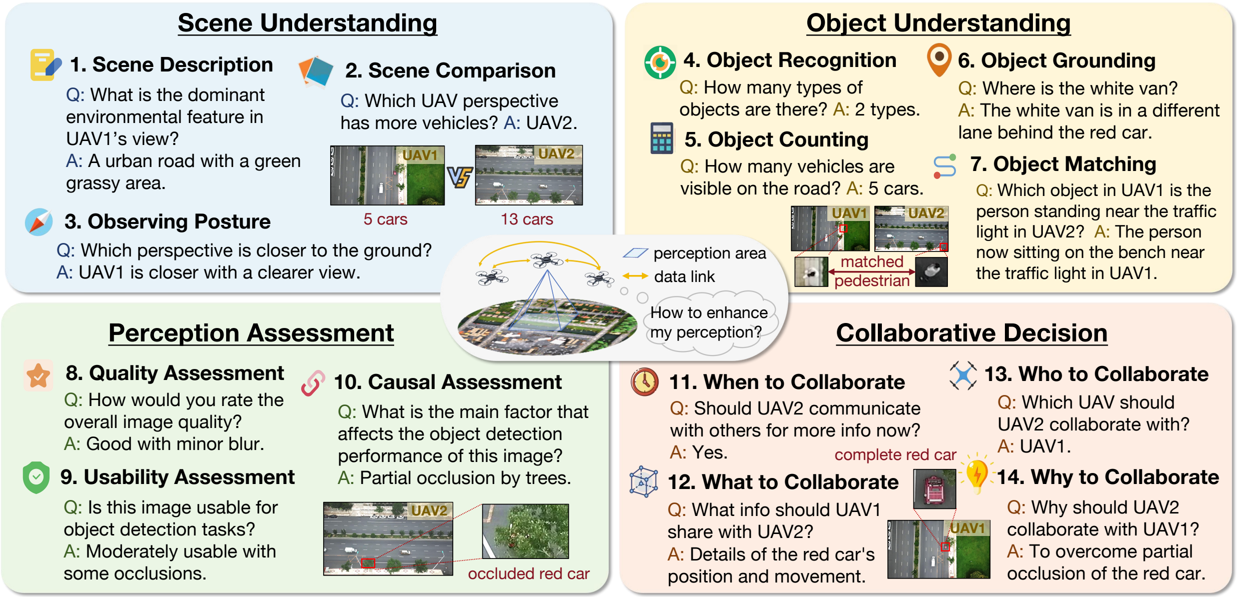
Table 1: Comparison of the proposed and popular benchmarks for collaborative perception.

Accordingly, we propose **AirCopBench**, a novel benchmark designed to evaluate MLLMs in multi-UAV collaborative perception under various challenging perception conditions. Specifically, we introduce an innovative task set comprising 14 tasks across 4 dimensions from simultaneous multi-view images. Our dataset includes challenging data from both simulators (Gao et al. 2024) and real world (Liu et al. 2023), collected from diverse UAV groups in representative degraded scenarios. Besides traditional object-level labeling for perception tasks, we incorporate event-level labeling by human annotators for specific cooperation events. Then, we develop a general pipeline to generate high-quality visual question answering (VQA) pairs using model-based, rule-based, and human-based approaches, followed by stringent quality control measures, including standard review, blind filtering, and human refinement. Finally, we evaluate our dataset on popular MLLMs in zero-shot settings, including both proprietary and open-source models, and conduct supervised fine-tuning (SFT) on Qwen2.5-series (Bai et al. 2025a) and LLaVA-NeXT series (Liu et al. 2024) to validate the dataset's effectiveness. The results show that current MLLMs struggle with multi-view collaborative perception, exhibiting inconsistent performance across different task types.

Compared to existing aerial collaborative perception benchmarks, key features of AirCopBench include: 1) Semantic VQA pairs to evaluate MLLMs' ability in perception assessment and collaborative reasoning; 2) Various challenging perception degradation in real-world scenarios, such as occlusion, shadows, motion blur, noise, data loss, long-range detection, complex background, etc; 3) Multiple modalities, including RGB images, text, and point clouds, to support perception across diverse data types; 4) Diverse target categories, covering drones, pedestrians, vehicles, and bicycles; 5) Embodied multi-UAV collaboration from first-view, role-based reasoning for human-like decision-making.

Overall, the novelty of this research lies in creating **the first benchmark for semantic embodied aerial collaborative perception considering various challenging perception degradation**. Our contributions are fourfold:

- We introduce a novel task set with 4 categories and 14 tasks to evaluate MLLMs' abilities in scene and object understanding, perception assessment, and collaborative decision-making using multi-view images.
- We create an aerial collaborative perception dataset with over 2.9k+ multi-view images from UAV groups of varying sizes, annotated for collaborative events, and focused on diverse perception degradations.
- We construct 14.6k+ VQA pairs from collaborative UAV



1. Scene Description — Q: What is the dominant environmental feature in UAV1's view? A: A urban road with a green grassy area.

2. Scene Comparison — Q: Which UAV perspective has more vehicles? A: UAV2.

3. Observing Posture — Q: Which perspective is closer to the ground? A: UAV1 is closer with a clearer view.

4. Object Recognition — Q: How many types of objects are there? A: 2 types.

5. Object Counting — Q: How many vehicles are visible on the road? A: 5 cars.

6. Object Grounding — Q: Where is the white van? A: The white van is in a different lane behind the red car.

7. Object Matching — Q: Which object in UAV1 is the person standing near the traffic light in UAV2? A: The person now sitting on the bench near the traffic light in UAV1.

8. Quality Assessment — Q: How would you rate the overall image quality? A: Good with minor blur.

9. Usability Assessment — Q: Is this image usable for object detection tasks? A: Moderately usable with some occlusions.

10. Causal Assessment — Q: What is the main factor that affects the object detection performance of this image? A: Partial occlusion by trees.

11. When to Collaborate — Q: Should UAV2 communicate with others for more info now? A: Yes.

12. What to Collaborate — Q: What info should UAV1 share with UAV2? A: Details of the red car's position and movement.

13. Who to Collaborate — Q: Which UAV should UAV2 collaborate with? A: UAV1.

14. Why to Collaborate — Q: Why should UAV2 collaborate with UAV1? A: To overcome partial occlusion of the red car.

Figure 2: AirCopBench includes 14 task types across 4 evaluation dimensions: Scene Understanding, Object Understanding, Perception Assessment, and Collaborative Decision. This categorization facilitates a systematic evaluation of MLLMs, from image understanding and quality analysis to multi-UAV information exchange for improved collaborative embodied perception.

observations using both real and simulated data, and design an extensible benchmark pipeline applicable to other multi-view embodied settings.

- We evaluate 40 MLLMs and investigate the relationship between embodied collaborative perception tasks and Sim-to-Real potential. We also fine-tune 4 models to demonstrate the effectiveness of our dataset.

2 AirCopBench

AirCopBench is a high-quality aerial perception benchmark with 14 tasks across 4 dimensions, including challenging perception degradation scenarios. Based on quantitative multi-view images from various drone groups, it evaluates MLLMs' embodied collaborative perception capabilities. This section outlines the task definitions, generation pipeline, and statistical properties of the benchmark.

2.1 Benchmark Tasks

AirCopBench includes four key task dimensions: Scene Understanding, Object Understanding, Perception Assessment, and Collaborative Decision. Each category is further divided into sub-tasks to enable a more detailed evaluation of MLLMs' capabilities, as shown in Fig. 2.

Scene Understanding refers to interpreting and understanding scenes from multi-view images (Task 1-3 in Fig. 2). Specifically, Scene Description identifies objects, relationships, and context in one scene, while Scene Comparison reveals similarities and differences between images from different views (Zhou et al. 2025). Observing Posture analyzes the relative distances and directions between observers and the scene, crucial for simulating human-like observations in embodied aerial perception tasks.

Object Understanding focuses on analyzing objects in terms of identity, quantity, and spatial relationships (Tasks 4-7 in Fig. 2). Particularly, Object Recognition identifies object types, Object Counting determines object quantity for specific types, Object Grounding assesses the spatial positioning between objects, and Object Matching finds corresponding objects across different views (Zhou et al. 2025).

Perception Assessment evaluates image quality and usability for specific perception tasks (Task 8-10 in Fig. 2). In detail, Quality Assessment rates image by objectively checking its clarity and resolution, while Usability Assessment subjectively determines its effectiveness for object detection (Wang et al. 2025c). Causal Assessment traces the reasons for poor-quality or unsuitable images, analyzing whether degradation is due to sensor issues, moving targets, or environmental factors. This helps UAVs better understand their perception state and supports information exchange with other observing UAVs.

Collaborative Decision determines the need for information exchange among UAVs (Task 11-14 in Fig. 2). This task design aims to enhance collaboration efficiency between UAVs by reducing communication and computational costs of the multi-UAV system. Specifically, When to Collaborate identifies scenarios where collaboration is essential for current UAV, while What to Collaborate determines the information that should be shared between UAVs. Who to Collaborate assesses which UAVs are best suited for collaboration, and Why to Collaborate explores the motivations behind information exchange between UAVs (Chen et al. 2024).

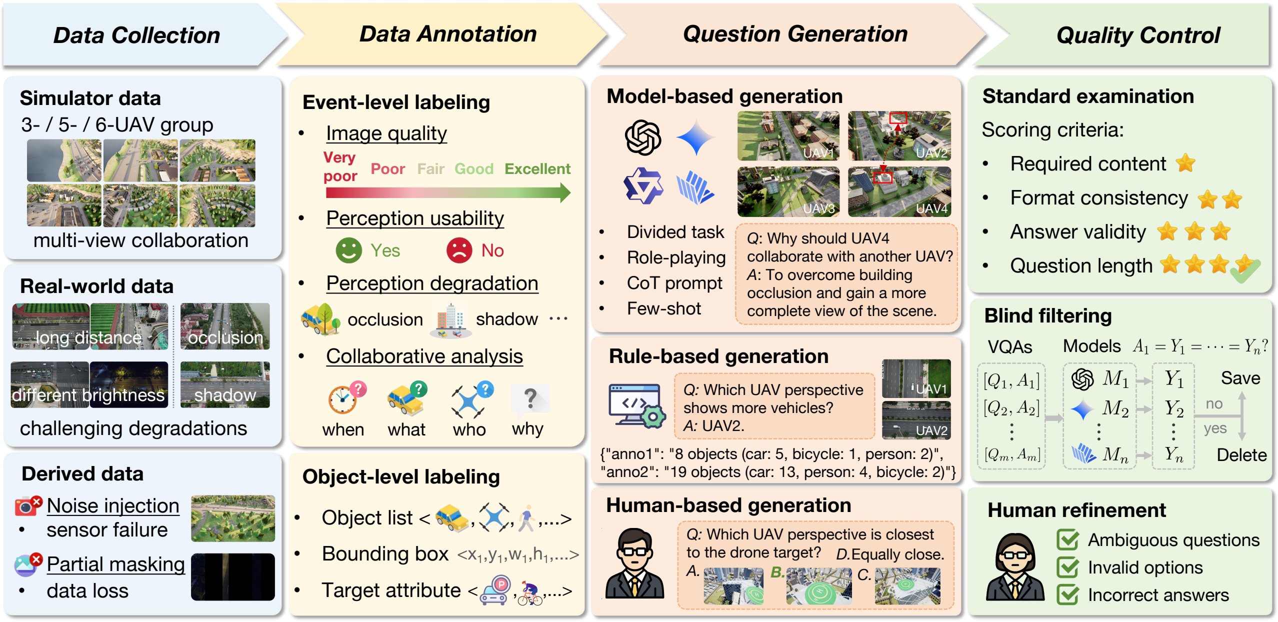


Figure 3: AirCopBench generation pipeline includes 4 main steps: Data Collection, Data Annotation, Question Generation, and Quality Control. This systematic approach ensures the validity and high quality of the generated dataset.

2.2 Benchmark Generation

The benchmark generation pipeline consists of four main steps: Data Collection, Data Annotation, Question Generation, and Quality Control, as illustrated in Fig. 3.

Data Collection. To effectively evaluate MLLMs in challenging scenarios, high-quality data with realistic perception degradations, such as occlusion, shadow, noise, lighting imbalance, data loss, motion blur, long distance, and out-of-field of view (FoV), is essential. Our data consists of:

Simulator Data. Multimodal collaborative perception data, including RGB images and point clouds, is collected from co-simulated Carla (Dosovitskiy et al. 2017) and AirSim (Shah et al. 2017). This includes challenging ground vehicle target perception from 5-6 UAVs in existing datasets, Coperception-UAV (Hu et al. 2022) and AeroCollab3D (Tian et al. 2024), and drone target observation from 3 UAVs collected from EmbodiedCity (Gao et al. 2024). The data spans 20 different map scenarios, focusing on those inducing key perceptual degradation issues, as detailed in *Appendix*.

Real-world Data. To enhance data credibility, we select representative images from existing real-world dataset MDMT (Liu et al. 2023), where 2 UAVs cooperatively perceive ground targets like vehicles, pedestrians, and bicycles. These scenarios include various perception challenges such as motion blur, varying lighting (day/night), distant small targets, and obstacle occlusions.

Derived Data. To expand the dataset and introduce more perceptual degradation, we apply two image post-processing techniques: a) Noise Injection adds random noises (Gaussian, salt-and-pepper, impulse, Poisson) to simulate sensor failures; b) Partial Masking blocks part of the image to simulate visual information loss. These techniques simulate perception degradation scenarios, such as low signal-to-noise ratio (SNR) and data loss due to sensor damage or failures, that are hard to capture in simulators or real environments.

Data Annotation. Rich, rational annotations tailored for aerial collaborative perception improve dataset accuracy and realism. In this work, we focus on both event-level labels, capturing inter-UAV collaboration, and object-level labels, ensuring 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.

Question Generation. Given diverse complexity and requirements for different collaborative perception tasks, we design three approaches for question generation:

Model-based Generation. This method uses powerful MLLMs, including GPT-4o (Hurst et al. 2024) and Qwen-VL-Max-latest (Bai et al. 2023), to efficiently generate high-quality VQA pairs for each task. We employ four prompting techniques: a) Task Decomposition breaks large tasks into smaller, focused tasks for better model understanding and

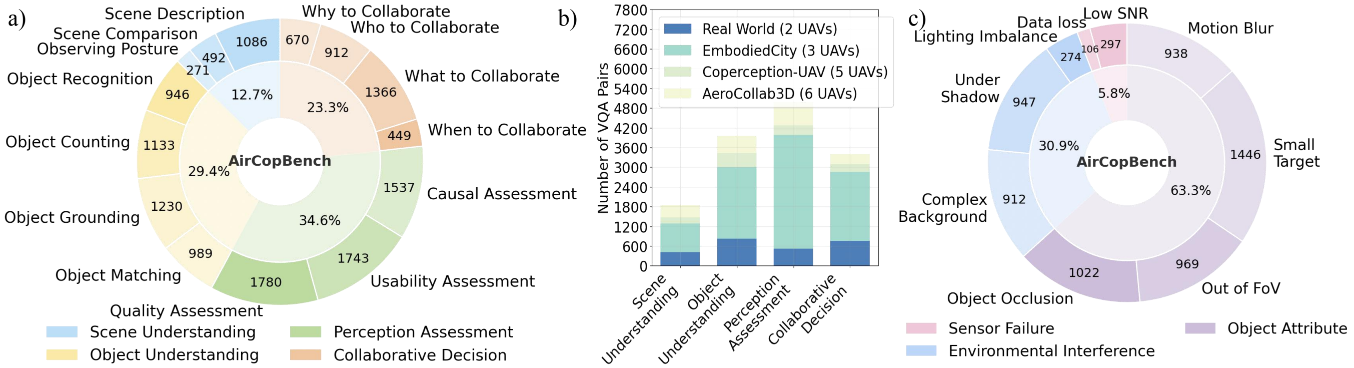


Figure 4: **Statistical Overview of AirCopBench.** a) Distribution of VQA pairs across 14 task types. b) Distribution of VQA pairs from various data sources with different numbers of observing UAV groups. c) Distribution of images featuring diverse perception degradation types.

easier debugging; b) Role-playing Settings prompts models to generate questions from the perspective of different UAV agents in embodied perception tasks; c) Chain-of-Thought (CoT) Prompting uses multi-step reasoning to generate complex questions requiring deep visual understanding; d) Few-shot Learning leverages templates and examples for improved task understanding and generalization.

Rule-based Generation. This approach uses predefined rules and logic to generate structured questions. Tasks like Object Counting and Usability Assessment rely on rule-based methods, where questions are directly based on dataset annotations. These simple, deterministic rules make rule-based generation ideal for tasks requiring consistent, structured questions without complex reasoning or context.

Human-based Generation. This method supports tasks like Observing Posture and Object Matching, which require complex multi-image reasoning and spatial understanding. Expert annotations ensure subtle visual cues, realistic dynamics, and logical nuances are well captured, yielding high-quality, context-rich questions.

Quality Control. To ensure the reliability and effectiveness of our dataset for training and evaluating models on collaborative perception, we implement three key quality control measures: Standard Examination, Blind Filtering, and Human Refinement.

Standard Examination. This measure evaluates the quality of generated VQA pairs based on four criteria: a) Required Content ensures all necessary information is included, flagging incomplete pairs for revision; b) Format Consistency maintains uniform structure, wording, and presentation; c) Answer Validity checks the correctness and relevance of options, filtering out incorrect ones and ensuring the correct option is present; d) Question Length ensures questions are sufficiently detailed to avoid ambiguity.

Blind Filtering. This measure aims to remove questions answerable through common sense by using n MLLMs to predict answers without the multi-view image input. If all MLLMs answer correctly, the question is removed, filtering out invalid questions that can be answered solely through general perception knowledge (Zhao et al. 2025a).

Human Refinement. This step further ensures human modification of generated questions with issues such as: a) Ambiguous Questions that are unclear, poorly framed, or irrelevant to the perception task; b) Invalid Options that are missing correct answers, contain duplicates or mismatched content; c) Incorrect Answers that are missing, contain multiple selections when only one is required, or are incorrectly chosen. The entire human refinement process takes over 800 hours. Specific modified examples are provided in *Appendix*.

2.3 Benchmark Statistics

AirCopBench comprises 2,920 simultaneous multi-view images with various challenging perception degradation collected from real world and simulators, as shown in Fig. 4b&c. The dataset includes 14,610 questions, including both basic tasks like scene and object understanding, as well as advanced tasks like perception assessment and collaborative decision, as detailed in Fig. 4a.

3 Experiments

This section introduces the evaluated models and protocols, evaluates 40 mainstream MLLMs on AirCopBench, and summarizes results across model types. It also analyzes task correlations, sim-to-real potential, and error cases.

3.1 Experimental Setup

Evaluation Metric. We integrate AirCopBench evaluation into the VLMEvalKit (Duan et al. 2024) framework, enabling capability assessments of new MLLMs. Specifically, we measure MLLM accuracy for each task type using question-answering accuracy.

Baselines. We conduct zero-shot evaluations of 40 MLLMs on AirCopBench in Tab. 2. The baselines include both proprietary models, such as GPT-4o (Hurst et al. 2024), Gemini-2.5-Pro (Comanici et al. 2025), Claude-Sonnet-4 (Anthropic 2025), and Qwen-Max-VL (Bai et al. 2023), as well as open-source models capable of multi-image input, including LLaVA series (Liu et al. 2024), Qwen-VL series (Bai et al. 2025a; Wang et al. 2024b) and Phi series (Abdin et al. 2024a,b), among others. Additional evaluated models include (Chen et al. 2025; Lu et al. 2023; Yao et al. 2024;

			Scene Understanding			Object Undstanding				Perception Assessment			Collaborative Decision			
Method	Rank	Avg.	*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

Table 2: Results on AirCopBench for existing various MLLMs on 14 task types across 4 evaluation dimensions. The best-performing model in each category is highlighted **in-bold**, while the second-best is underlined. 24 out of 40 models and 3 out of 4 fine-tuned models for demonstration in the main text; additional results are provided in the *Appendix*.

Team et al. 2025; Yue et al. 2025; Ye et al. 2024; Shen et al. 2025; Lu et al. 2024; Wang et al. 2025a; Beyer et al. 2024; Zhu et al. 2025).

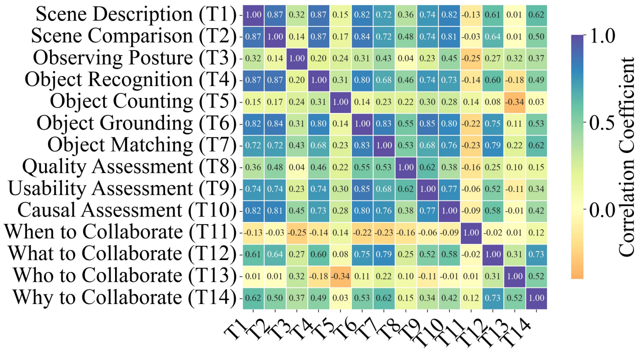
3.2 Model Comparison

From the quantitative results shown in Tab. 2, we have the following findings:

- **AirCopBench poses significant challenges to all MLLMs.** Both proprietary and open-source models perform poorly on aerial collaborative perception tasks, with the best model, Ovis-16B (Lu et al. 2024), achieving only 59.17% accuracy. This highlights the importance of AirCopBench in revealing the insufficient development of embodied perception and collaborative decision-making abilities in current MLLMs.
- **Leading open-source MLLMs match or exceed the best proprietary models on AirCopBench.** Models like Ovis2-16B (Lu et al. 2024) and Kimi-VL-A3B-Thinking (Team et al. 2025) match or slightly surpass top proprietary systems. However, the overall performance gap remains modest, and both open-source and proprietary MLLMs still struggle with multi-image understanding, underscoring the need for further progress in multi-image reasoning.
- **MLLMs demonstrate a clear bias across various question categories.** Current MLLMs perform well on tasks like Scene Description and Scene Comparison, but struggle with categories requiring domain knowledge and goal-oriented reasoning, such as Usability Assessment and When to Collaborate. This indicates a need for significant improvements in embodied perception assessment and adaptive collaboration decision-making.

3.3 Correlation Analysis

To explore the relationships between tasks and their required cognitive abilities, we compute pairwise correlations



	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
Scene Description (T1)	1.00	0.87	0.32	0.87	0.15	0.82	0.72	0.36	0.74	0.82	-0.13	0.61	0.01	0.62
Scene Comparison (T2)	0.87	1.00	0.14	0.87	0.17	0.84	0.72	0.48	0.74	0.81	-0.03	0.64	0.01	0.50
Observing Posture (T3)	0.32	0.14	1.00	0.20	0.24	0.31	0.43	0.04	0.23	0.45	-0.25	0.27	0.32	0.37
Object Recognition (T4)	0.87	0.87	0.20	1.00	0.31	0.80	0.68	0.46	0.74	0.73	-0.14	0.60	-0.18	0.49
Object Counting (T5)	0.15	0.17	0.24	0.31	1.00	0.14	0.23	0.22	0.30	0.28	0.14	0.08	-0.34	0.03
Object Grounding (T6)	0.82	0.84	0.31	0.80	0.14	1.00	0.83	0.55	0.85	0.80	-0.22	0.75	0.11	0.53
Object Matching (T7)	0.72	0.72	0.43	0.68	0.23	0.83	1.00	0.53	0.68	0.76	-0.23	0.79	0.22	0.62
Quality Assessment (T8)	0.36	0.48	0.04	0.46	0.22	0.55	0.53	1.00	0.62	0.38	-0.16	0.25	0.10	0.15
Usability Assessment (T9)	0.74	0.74	0.23	0.74	0.30	0.85	0.68	0.62	1.00	0.77	-0.06	0.52	-0.11	0.34
Causal Assessment (T10)	0.82	0.81	0.45	0.73	0.28	0.80	0.76	0.38	0.77	1.00	-0.09	0.58	-0.01	0.42
When to Collaborate (T11)	-0.13	-0.03	-0.25	-0.14	0.14	-0.22	-0.23	-0.16	-0.06	-0.09	1.00	-0.02	0.01	0.12
What to Collaborate (T12)	0.61	0.64	0.27	0.60	0.08	0.75	0.79	0.25	0.52	0.58	-0.02	1.00	0.31	0.73
Who to Collaborate (T13)	0.01	0.01	0.32	-0.18	-0.34	0.11	0.22	0.10	-0.11	-0.01	0.01	0.31	1.00	0.52
Why to Collaborate (T14)	0.62	0.50	0.37	0.49	0.03	0.53	0.62	0.15	0.34	0.42	0.12	0.73	0.52	1.00

Figure 5: Correlation coefficients of MLLMs' performance across all tasks, with higher values indicating greater similarity in the cognitive abilities required by the two tasks.

of MLLMs' accuracy (Tab. 1) on each task. Similar performance across models on two tasks suggests shared cognitive abilities. From Fig. 5, we draw the following conclusions:

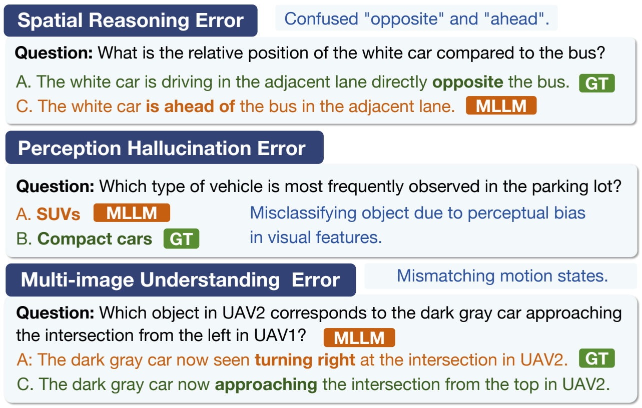
- **Causal Assessment exhibits a high correlation with almost all other tasks.** It suggests that understanding and inferring the causality of perception degradation is crucial for various cognitive processes. This finding implies that causal assessment could be a key factor in the development of embodied cognition in collaborative perception.
- **Object Matching has a high correlation with both Object Recognition and Object Grounding tasks.** This observation aligns with prior knowledge that effective multi-view image understanding relies on strong single-image perceptual capabilities.
- **Quality Assessment exhibits moderate correlations with multiple tasks.** This suggests that quality evaluation requires integrating Scene Understanding, Object Recognition, and Object Matching abilities, making it a highly composite decision-making task.

3.4 Supervised Fine-Tuning

We fine-tune two representative MLLMs, Qwen-2.5-VL (7B and 3B) (Bai et al. 2025a) and LLaVA-NeXT-13B (Liu et al. 2024), using our curated instruction dataset within the LLaMA-Factory framework (Zheng et al. 2024), with default hyperparameters for three epochs. As shown in Tab. 2, Qwen-2.5-VL-7B achieves 74.30% accuracy (+26.97), Qwen-2.5-VL-3B reaches 66.44% (+19.11), and LLaVA-NeXT-13B improves to 57.61% (+19.30), demonstrating that domain-specific SFT substantially enhances MLLM performance in collaborative perception.

3.5 Sim-to-Real

We evaluate the generalization of models trained on simulated data to real-world UAV imagery. As shown in Tab. 2, we compare the open-source Qwen2.5-VL-7B (Bai et al. 2025a) and our AirCop-7B model fine-tuned on simulator data. Qwen2.5-VL-7B achieves 47.77% overall accuracy, performing well in Object Recognition (83.33%) but poorly in Posture (11.11%) and Collaboration (40%). After fine-tuning, AirCop-7B reaches 67.41% (+19.64), with notable gains in Object Grounding (77.78%, +50.00), Object Counting (88.89%, +38.89), and Usability Assessment (94.44%, +53.20). These results confirm that simulator-based fine-tuning significantly enhances sim-to-real transfer, improving generalization to real UAV imagery and validating the effectiveness of our simulated dataset.



Question: What is the relative position of the white car compared to the bus?

A. The white car is driving in the adjacent lane directly **opposite** the bus. GT

C. The white car **is ahead of** the bus in the adjacent lane. MLLM

Question: Which type of vehicle is most frequently observed in the parking lot?

A. **SUVs** MLLM

B. **Compact cars** GT

Misclassifying object due to perceptual bias in visual features.

Question: Which object in UAV2 corresponds to the dark gray car approaching the intersection from the left in UAV1? MLLM

A: The dark gray car now seen **turning right** at the intersection in UAV2. GT

C. The dark gray car now **approaching** the intersection from the top in UAV2.

Figure 6: Examples of three common errors in MLLM reasoning during aerial collaborative perception tasks.

3.6 Error Analysis

The errors in MLLM reasoning primarily stem from three causes, as shown in Fig. 6:

- **Perception hallucination errors.** MLLMs misidentify or fail to recognize objects due to perception hallucination, where visual input is incorrectly processed, leading to false or exaggerated perceptions of objects.
- **Spatial reasoning errors.** MLLMs struggle with spatial reasoning, causing errors in interpreting object relationships, such as incorrect positioning or orientation of objects in space (Zhang et al. 2025a; Zha et al. 2025).
- **Multi-image understanding errors.** MLLMs face challenges in multi-image understanding, resulting in errors when comparing, matching, or integrating information across images, leading to incorrect logical inferences.

Perceptual and spatial errors hinder Object Recognition and Grounding, while multi-image reasoning limits Scene Comparison, Object Matching, and Collaborative Decision; other tasks show varied failure modes (see *Appendix*).

4 Conclusion

In this work, we propose AirCopBench, a benchmark for multi-UAV collaborative perception with challenging degradation, featuring 2.9k+ multi-view images and 14.6k+ questions. We evaluate 40 popular MLLMs on scene understanding, object understanding, perception assessment, and collaborative decision. The best-performing models achieve only 59.17% accuracy, highlighting challenges in multi-view collaborative perception. Fine-tuned MLLMs with simulated data improves their performance on real-world tasks. Future work will extend to more diverse real-world data and explore efficient MLLM architectures for practical multi-UAV deployment (Ren et al. 2023).

Acknowledgments

This paper was supported by the Natural Science Foundation of China under Grant 62371269, Shenzhen Low-Altitude Airspace Strategic Program Portfolio Z253061 and Meituan Academy of Robotics Shenzhen. Sponsored by Tsinghua University-Toyota Research Center.

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Anthropic. 2025. Claude 3.7 Sonnet "Extended Thinking" System Card. Technical report.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; et al. 2025a. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, T.; Fan, Y.; Qiu, J.; Sun, F.; Song, J.; Han, J.; Liu, Z.; He, C.; Zhang, W.; and Yuan, B. 2025b. Hallucination at a Glance: Controlled Visual Edits and Fine-Grained Multimodal Learning. *arXiv preprint arXiv:2506.07227*.

Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Chen, X.; Xiao, Z.; Cheng, Y.; Hsia, C.-C.; Wang, H.; Xu, J.; Xu, S.; Dang, F.; Zhang, X.-P.; Liu, Y.; et al. 2024. Soscheduler: Toward proactive and adaptive wildfire suppression via multi-uav collaborative scheduling. *IEEE Internet of Things Journal*, 11(14): 24858–24871.

Cheng, Z.; Xu, B.; Gong, L.; Song, Z.; Zhou, T.; Zhong, S.; Ren, S.; Chen, M.; Meng, X.; Zhang, Y.; et al. 2025. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11198–11201.

Gao, C.; Zhao, B.; Zhang, W.; Mao, J.; Zhang, J.; Zheng, Z.; Man, F.; Fang, J.; Zhou, Z.; Cui, J.; et al. 2024. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kil, J.; Mai, Z.; Lee, J.; Chowdhury, A.; Wang, Z.; Cheng, K.; Wang, L.; Liu, Y.; and Chao, W.-L. H. 2024. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems*, 37: 28798–28827.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, Z.; Shang, Y.; Li, T.; Chen, G.; Wang, Y.; Hu, Q.; and Zhu, P. 2023. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. *IEEE Transactions on Multimedia*, 25: 1462–1476.

Lu, P.; Peng, B.; Cheng, H.; et al. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36: 43447–43478.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Ren, J.; Xu, Y.; Li, Z.; Hong, C.; Zhang, X.-P.; and Chen, X. 2023. Scheduling uav swarm with attention-based graph reinforcement learning for ground-to-air heterogeneous data communication. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, 670–675.

Saini, N.; Dubey, A.; Das, D.; and Chattopadhyay, C. 2025. Advancing open-set object detection in remote sensing using multimodal large language model. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 451–458.

Shah, S.; Dey, D.; Lovett, C.; and Kapoor, A. 2017. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics: Results of the 11th international conference*, 621–635. Springer.

Shen, W.; Pei, J.; Peng, Y.; Song, X.; Liu, Y.; Peng, J.; Sun, H.; Hao, Y.; Wang, P.; and Zhou, Y. 2025. Skywork-R1V3 Technical Report. *arXiv preprint arXiv:2507.06167*.

Sunderraman, R.; Ji, J. S.; et al. 2024. UAV3D: A Large-scale 3D Perception Benchmark for Unmanned Aerial Vehicles. *Advances in Neural Information Processing Systems*, 37: 55425–55442.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Tian, P.; Wang, Z.; Cheng, P.; Wang, Y.; Wang, Z.; et al. 2024. Ucdnet: Multi-uav collaborative 3d object detection network by reliable feature mapping. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, F.; Fu, X.; Huang, J. Y.; Li, Z.; Liu, Q.; Liu, X.; Ma, M. D.; Xu, N.; Zhou, W.; Zhang, K.; et al. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Wang, G.-H.; Zhao, S.; Zhang, X.; Cao, L.; Zhan, P.; Duan, L.; Lu, S.; Fu, M.; Chen, X.; Zhao, J.; et al. 2025a. Ovis-U1 Technical Report. *arXiv preprint arXiv:2506.23044*.

Wang, H.; Chen, X.; Cheng, Y.; Wu, C.; Dang, F.; and Chen, X. 2022. H-swarmloc: Efficient scheduling for localization of heterogeneous mav swarm with deep reinforcement learning. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 1148–1154.

Wang, H.; Xu, J.; Luo, X.; Chen, X.; Zhang, T.; Duan, R.; Liu, Y.; and Chen, X. 2025b. Ultra-high-frequency harmony: mmwave radar and event camera orchestrate accurate drone landing. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, 15–29.

Wang, H.; Xu, J.; Luo, X.; Zhang, T.; Chen, X.; Duan, R.; Chen, J.; Liu, Y.; Zheng, J.; Hong, W.; et al. 2025c. mmE-Loc: Facilitating Accurate Drone Landing with Ultra-High-Frequency Localization. *arXiv preprint arXiv:2507.09469*.

Wang, H.; Xu, J.; Zhao, C.; Cheng, Y.; Chen, X.; Hong, C.; Zhang, X.-P.; Liu, Y.; and Chen, X. 2025d. Aerial Shepherds: Enabling Hierarchical Localization in Heterogeneous MAV Swarms. *IEEE Transactions on Mobile Computing*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Cheng, P.; Chen, M.; et al. 2024c. Drones help drones: A collaborative framework for multi-drone object trajectory prediction and beyond. *Advances in Neural Information Processing Systems*, 37: 64604–64628.

Xiao, J.; Sun, Y.; Shao, Y.; Gan, B.; Liu, R.; Wu, Y.; Guan, W.; and Deng, X. 2025. Uav-on: A benchmark for open-world object goal navigation with aerial agents. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 13023–13029.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13040–13051.

Yeh, C.-H.; Wang, C.; Tong, S.; Cheng, T.-Y.; Wang, R.; Chu, T.; Zhai, Y.; Chen, Y.; Gao, S.; and Ma, Y. 2025. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*.

Yue, X. L.-C. T. Z.; Lin, Z.; Song, Y.-H.; Wang, W.; Ren, S.-Q.; Gu, S.; Li, S.-G.; Li, P.; Zhao, L.; Li, L.; Bao, K.; Tian, H.; Zhang, H.; Wang, G.; Zhu, D.; Cici; He, C.; Ye, B.; Shen, B.; Zhang, Z.; Jiang, Z.-A.; Zheng, Z.; Song, Z.; et al. 2025. MiMo-VL Technical Report. *ArXiv*, abs/2506.03569.

Zha, J.; Fan, Y.; Yang, X.; Gao, C.; and Chen, X. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*.

Zha, J.; Zhou, N.; Liu, Z.; Sun, T.; and Chen, X. 2024. Diffusion-based filter for fast and accurate collaborative tracking with low data transmission. *Authorea Preprints*.

Zhang, W.; Peng, R.; Gao, C.; Fang, J.; Zeng, X.; Li, K.; Wang, Z.; Cui, J.; Wang, X.; Chen, X.; et al. 2025a. The Point, the Vision and the Text: Does Point Cloud Boost Spatial Reasoning of Large Language Models? *arXiv preprint arXiv:2504.04540*.

Zhang, W.; Zhou, Z.; Zeng, X.; Xuchen, L.; Fang, J.; Gao, C.; Cui, J.; Li, Y.; Chen, X.; and Zhang, X.-P. 2025b. Open3D-VQA: A Benchmark for Embodied Spatial Concept Reasoning with Multimodal Large Language Model in Open Space. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12784–12791.

Zhao, B.; Fang, J.; Dai, Z.; et al. 2025a. UrbanVideo-Bench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 32400–32423. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Zhao, B.; Wang, Z.; Fang, J.; Gao, C.; Man, F.; Cui, J.; Wang, X.; Chen, X.; Li, Y.; and Zhu, W. 2025b. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 11071–11080.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Bangkok, Thailand: Association for Computational Linguistics.

Zhou, B.; Yang, H.; Chen, D.; Ye, J.; Bai, T.; Yu, J.; Zhang, S.; Lin, D.; et al. 2025. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10707–10715.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.