MM-UAVBENCH: How Well Do Multimodal Large Language Models See, Think, and Plan in Low-Altitude UAV Scenarios?

Shiqi Dai[1*] Zizhi Ma[2*] Zhicong Luo[3] Xuesong Yang[4] Yibin Huang[5] Wanyue Zhang[4]
Chi Chen[1†] Zonghao Guo[1] Wang Xu[1] Yufei Sun[2] Maosong Sun[1†]

[1]Tsinghua University [2]Nankai University [3]Northwest Polytechnical University
[4]Chinese Academy of Sciences [5]Harbin Institute of Technology

{daisq99@gmail.com, chenchithu@gmail.com}

Project Page Evaluation Code MM-UAVBENCH

Abstract

While Multimodal Large Language Models (MLLMs) have exhibited remarkable general intelligence across diverse domains, their potential in low-altitude applications dominated by Unmanned Aerial Vehicles (UAVs) remains largely underexplored. Existing MLLM benchmarks rarely cover the unique challenges of low-altitude scenarios, while UAV-related evaluations mainly focus on specific tasks such as localization or navigation, without a unified evaluation of MLLMs' general intelligence. To bridge this gap, we present MM-UAVBENCH*, a comprehensive benchmark that systematically evaluates MLLMs across three core capability dimensions—perception, cognition, and planning—in low-altitude UAV scenarios.* MM-UAVBENCH *comprises 19 sub-tasks with over 5.7K manually annotated questions, all derived from real-world UAV data collected from public datasets. Extensive experiments on 16 open-source and proprietary MLLMs reveal that current models struggle to adapt to the complex visual and cognitive demands of low-altitude scenarios. Our analyses further uncover critical bottlenecks such as spatial bias and multi-view understanding that hinder the effective deployment of MLLMs in UAV scenarios. We hope* MM-UAVBENCH *will foster future research on robust and reliable MLLMs for real-world UAV intelligence.*

1. Introduction

With the rapid progress of Multimodal Large Language Models (MLLMs), their capabilities have become increasingly comprehensive [1, 39, 42]. Such integrated intelligence is especially appealing for Unmanned Aerial Vehicles (UAVs), which are evolving from passive sensing platforms into autonomous edge agents in complex low-altitude environments. Integrating MLLMs into UAVs can elevate their intelligence from basic perception-such as object detection [28] and tracking [46]—to cognitive reasoning and task planning, marking a key step toward autonomous aerial intelligence in real-world missions.

Despite this potential, the fundamental abilities of MLLMs to operate or assist in complex low-altitude environments remain largely unevaluated. Most existing benchmarks for evaluating MLLMs focus on general image or video understanding in everyday scenes [9, 10, 15, 32], emphasizing static perception from ground-level or object-centric views. Even when low-altitude imagery occasionally appears in these datasets [32], it is not treated as a distinct evaluation domain. Meanwhile, several remote sensing benchmarks [5, 18, 31] assess MLLMs from satellite or aerial perspectives, but they mainly involve high-altitude top-down views with stable geometry and coarse spatial resolution. Consequently, none of these datasets capture the dynamic, near-ground, and multi-agent characteristics inherent to low-altitude UAV scenarios.

Recently, several studies have begun to investigate the applicability of MLLMs in UAV-related scenarios [2, 6, 17, 22, 28, 41]. However, most of these efforts primarily focus on traditional perception tasks such as object detection [2, 6, 43], referring grounding [28], counting [11, 35] and target tracking [24]. Another line of evaluation focuses on navigation and control-oriented tasks, including trajectory following and path planning for UAVs [34, 37, 41], aiming to assess models' low-level motion understanding or decision execution. Although these benchmarks contribute to evaluating UAV perception and navigation, they remain task-specific and lack a comprehensive assessment of MLLMs' higher-level abilities in realistic low-altitude environments.

In realistic UAV operations, intelligence involves more than recognizing objects or following trajectories—it de-

*Equal contribution.
†Corresponding author.

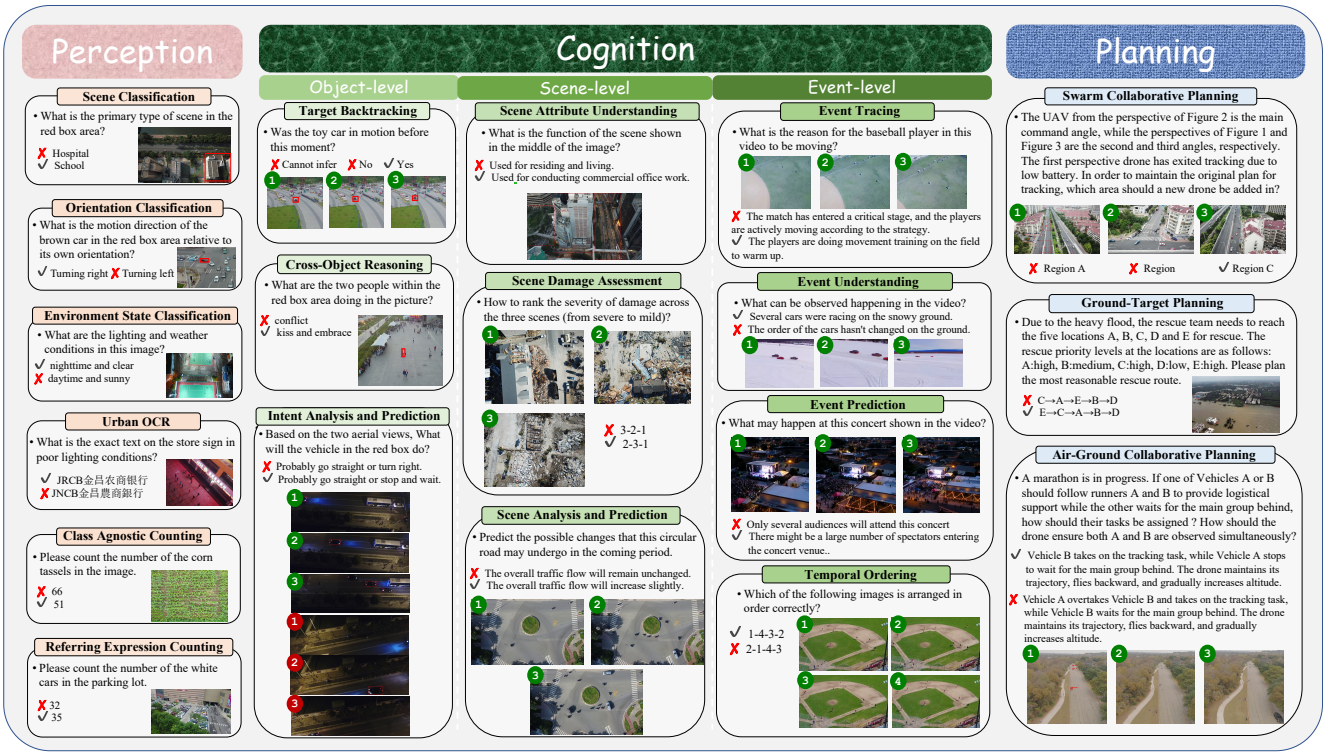


Figure 1. Overview of MM-UAVBENCH. MM-UAVBENCH consists of 19 tasks covering three core capability dimensions: Perception, Cognition, and Planning. Perception tasks assess basic visual understanding such as classification, OCR, and counting. Cognition tasks span three hierarchical levels—object-level, scene-level, and event-level—evaluating the model's ability to infer intentions, reason across objects, analyze scenes, understand events, and predict outcomes. Planning tasks assess UAV-specific decision making, including planning for single or multi-UAV systems, directing ground-target actions from an aerial perspective, and coordinating cooperative actions between aerial agents and ground participants. All examples shown are real UAV imagery, illustrating the diverse challenges present in low-altitude scenarios.

mands understanding what is happening within a scene, how multiple entities (including UAVs and ground targets) interact, and what strategic decisions should follow [30]. **A systematic evaluation benchmark is needed to measure how well MLLMs see, think, and plan in complex real-world UAV scenarios**.

In this work, we introduce **MM-UAVBENCH**, a comprehensive benchmark designed to evaluate the perception, cognition, and planning abilities of MLLMs in low-altitude UAV scenarios. MM-UAVBENCH provides a unified evaluation paradigm that reflects the hierarchical intelligence required for real-world aerial missions. It features three main characteristics:

- **Comprehensive Task Design.** MM-UAVBENCH includes 19 tasks across three key capability dimensions and incorporates UAV-specific considerations, including multi-level cognition (object, scene, and event) and planning that involves both aerial and ground agents, resulting in a comprehensive task design.
- **Diverse Real-World Scenarios.** Unlike previous benchmarks that focus on limited scenes or rely on simulated environments, MM-UAVBENCH is constructed from real UAV imagery collected across a wide range of scenarios, including but not limited to urban areas, agricultural fields, wildlife habitats, and emergency or disaster zones, enabling robust and generalizable evaluation.
- **High-quality Human Annotations.** All tasks are manually annotated to ensure both labeling quality and appropriate task difficulty. In addition, we provide multiple forms of detailed auxiliary annotations, such as bounding boxes for key entities, to support in-depth capability analysis.

To construct this benchmark, we collect real-world UAV videos and images from diverse data sources, encompassing 1549 video clips and 2873 images with an average resolution of 1622×1033. Using these data, we manually annotate 16 tasks, while the remaining 3 tasks are generated through rule-based transformation of manually annotated labels, resulting in 5702 multiple-choice QA annotations in total. We evaluate a broad set of MLLMs on MM-UAVBENCH and find that their perception capabilities in UAV scenarios remain limited, with even more pronounced deficiencies in cognition and planning tasks. Further analyses on object-scale sensitivity, spatial perception bias, multi-view understanding, and egocentric planning indicate that current MLLMs struggle to

Table 1. Overview of MM-UAVBENCH and comparison with representative existing benchmarks. ✅, ⊗, and ✗respectively denote datasets constructed from real imagery, purely simulation, or partially real data that include simulated components. The "Anno." column specifies the annotation method of each benchmark, including *Human* (purely human-labeled), *Auto* (fully automatic generation), and *Semi-Auto* (generated labels with human refinement). Scenario icons: Urban, Natural scenes, Wildlife, Disaster/Emergency, Agriculture.

Benchmark	Capability Types	#Tasks	Scenarios	Real Imagery	Anno.	#Source	#Test Instances
Remote Sensing Perspective							
VRSBench [18]	Perception / Cognition	31	–	✅	Semi-Auto	29.6K images	205.3K
XLRS-Bench [31]	Perception / Cognition	16	–	✅	Semi-Auto	1.4K images	45.9K
UAV Perspective							
UAVDT [6]	Object Detection and Tracking	3	Urban	✅	Human	80K images	841.5K
RefDrone [28]	Visual Grounding	1	Urban	✅	Semi-Auto	8.5K images	63.6K
UAV-Human [17]	Human Behavior Understanding	4	Urban	✅	Human	67.4K videos	86.0K
UAV-ON [37]	Visual-Language Navigation	1	Urban, Natural	⊗	Auto	1.2K targets	1.2K
OpenUAV [34]	Visual-Language Navigation	1	Urban, Natural	⊗	Auto	12.1K trajectories	12.1K
MME-RealWorld-MO [19]	High-resolution Understanding	6	Urban	✅	Human	1.6K images	2.2K
SkyAgent-Eval [38]	Embodied Capability	5	Urban	⊗	Human	67.4K videos	86.0K
UrbanVideo-Bench [41]	Embodied Capability	16	Urban	✗	Semi-Auto	1.5K videos	5.2K
MM-UAVBENCH (Ours)	**Comprehensive Per. / Cog. / Plan.**	**19**	Urban, Natural, Wildlife, Disaster/Emergency, Agriculture	✅	**Human**	**1.5K videos + 2.8K images**	**5.7K**

adapt to low-altitude UAV challenges, underscoring the need for UAV-tailored model designs for practical deployment. Our main contributions are summarized as follows:

- We present MM-UAVBENCH, a new and comprehensive benchmark for evaluating the perception, cognition, and planning capabilities of MLLMs across 19 tasks in low-altitude UAV scenarios.
- We construct MM-UAVBENCH from real-world UAV datasets with both manually annotated and rule-converted tasks, resulting in 5702 high-quality annotations that provide strong data authenticity and well-controlled task difficulty.
- We benchmark a series of MLLMs on MM-UAVBENCH and provide detailed analyses that expose critical limitations, highlighting the need for UAV-oriented MLLM designs for real-world deployment.

2. Related Work

2.1. General MLLM Benchmark

A wide range of benchmarks have been developed to evaluate the visual and reasoning abilities of MLLMs. General-purpose benchmarks such as MME [9], SEED-Bench [15], and VideoMME [10] offer broad assessments of object recognition, commonsense reasoning, and video understanding across everyday scenes. Recently, several works have also explored comprehensive MLLM evaluation in remote sensing, including VRSBench [18] and XLRS-Bench [31], but these operate primarily on high-altitude, top-down imagery with stable viewpoints. Despite their breadth, they do not address the distinct characteristics of low-altitude UAV scenarios, such as dynamic viewpoints, large scale variation, multi-entity interactions, and action-oriented decision making, and thus provide limited insight into the operational intelligence required for UAV missions.

2.2. Evaluation in Low-Altitude UAV Scenarios

Existing benchmarks for UAV scenarios mainly cover narrow and task-specific capabilities. Many perception-oriented datasets such as UAVDT [6] and RefDrone [28] focus on detection, tracking, or grounding in limited urban scenes. A second line of work, such as UAV-ON [37] and OpenUAV [34], evaluates visual-language navigation in simulator environments. Although these benchmarks introduce decision-oriented tasks, they rely heavily on synthetic scenes and address only ego-centric navigation. More recent embodied UAV evaluations, such as SkyAgent-Eval [38] and UrbanVideo-Bench [41], examine how MLLMs can assist UAVs in scene perception and flight planning. However, they still lack comprehensive assessments of UAV-perspective scene and event understanding, and their heavy reliance on simulator-generated data introduces potential sim-to-real gaps. Overall, existing UAV benchmarks offer limited capability coverage and restricted scenario diversity, providing only a partial view of the intelligence required for low-altitude UAV operations. In contrast, MM-UAVBENCH jointly evaluates perception, multi-level cognition, and multi-agent planning across diverse real-world UAV scenarios. A detailed comparison is provided in Table 1.

3. MM-UAVBENCH

In this section, we first introduce the hierarchical task design of MM-UAVBENCH. It spans 3 L1 catigories, 8 subcatigories, and 19 fine-grained tasks that are specifically tailored for UAV scenarios, as shown in Fig 2. Next, we describe the data collection process, question annotation procedures, and quality control guidelines, which greatly enhances the dataset's reliability and difficulty. Finally, we present a statistical overview of MM-UAVBENCH.

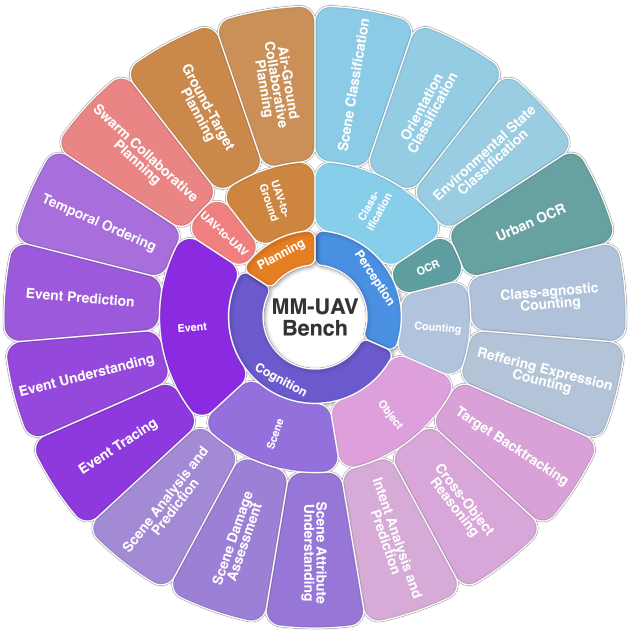


Figure 2. The task design of MM-UAVBENCH covers 3 high-level categories, 8 sub-catigories and 19 fine-grained tasks in MM-UAVBENCH.

3.1. Hierarchical Task Design

The diverse tasks in MM-UAVBENCH are carefully designed to comprehensively evaluate the capabilities of MLLMs, with each task associated with both L1 and L2 categories representing different levels of ability. The detailed definitions of each task under L2-level are listed in Appendix.

Perception. This dimension consists of three subcategories: classification, OCR, and counting. **1) Classification.** Identifying the category of objects or scenes in images. In UAV scenarios, such tasks include recognizing land-cover types (e.g., roads, buildings, farmlands) and transportation vehicles (e.g., cars, ships, airplanes). In particular, we annotate a large number of vehicle orientation classification tasks, which are crucial for road safety monitoring and trajectory prediction. **2) OCR (Optical Character Recognition).** Recognizing textual and symbolic information in images, mainly focusing on extracting information from road signs, markings, and traffic signals, which can support navigation and traffic management. **3) Counting.** Estimating the number of objects such as vehicles, people, or animals. In UAV scenarios, counting is valuable for traffic flow analysis, crowd density monitoring, and wild animal protection.

Cognition. Based on the reasoning target, cognition can be categorized into object-level, scene-level, and event-level reasoning. **1) Object-level.** Reasoning about the positions and behaviors of single or multiple target objects across past, present, and future spatial-temporal sequences, which supports the analysis of object trajectories, behavioral patterns, and anomalies. **2) Scene-level.** This includes three tasks: scene attribute understanding, scene damage assessment (e.g., fire, flood), and scene flow prediction, aimed at understanding the overall environmental state and its dynamic changes. **3) Event-level.** Reasoning about the causes, content, prediction, and temporal order of events, which helps UAVs identify events and anticipate their trends.

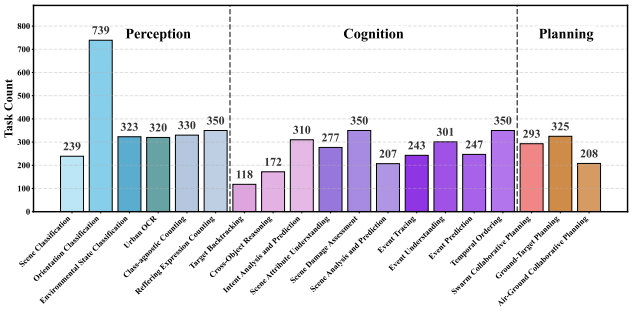
Planning. Based on the planning entities, planning can be categorized into two types: UAV-to-UAV planning and UAV-to-ground planning(including collaborative) planning.**1) UAV-to-UAV level.** For a small UAV group (e.g., three UAVs) executing joint missions, planning is conducted from two perspectives: task allocation and fault tolerance. Task allocation is based on the perspective of the UAV with the most comprehensive information (the command UAV), assigning roles and paths to each UAV to optimize overall group efficiency. Fault tolerance ensures that the group can still accomplish its mission even if individual UAVs fail or are disrupted. This capability is critical for tasks such as multi-UAV cooperative tracking, inspection, and search operations. **2) UAV-to-Ground level.** This level covers Ground-Target Planning and Air–Ground Collaborative Planning, where UAVs guide the movements of ground agents (e.g., rescue teams or vehicles) as well as their own trajectories based on environmental conditions and mission objectives, thereby enabling effective coordination between aerial and ground systems.

3.2. Dataset Construction

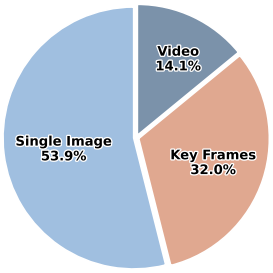
The curation of MM-UAVBENCH can be divided into three main components. First, we collect a large and diverse set of real UAV scenarios. Second, we adopt two annotation pipelines, manual labeling and rule-based conversion from existing datasets, while following a set of principles to ensure data quality. Finally, statistical analysis demonstrates the diversity and comprehensiveness of our benchmark.

3.2.1. Data Collection

We collect open-source datasets and conduct re-annotation to construct MM-UAVBENCH. The statistics of annotated datasets for each task are summarized in Appendix. These datasets not only encompass diverse environments such as urban and wilderness settings but also cover extreme scenarios including natural disasters (e.g., floods, wildfires) and human-induced incidents (e.g., violent events, traffic accidents). Moreover, they exhibit substantial diversity across temporal dimensions (day/night, seasonal variations, weather conditions) and geographical dimensions (countries, landscapes). The raw datasets we select follow two main criteria: 1. Data is collected by UAVs in the real world; 2. The datasets contain rich annotation which is beneficial for multiple-choice question generation. For video datasets, we



(a) Tasks distribution.



(b) Input modality.

Metric	Value
N_{video}	1549
N_{img}	2873
Avg. Res.	1622×1033
$N_{\text{bbox, man}}$	1267
$N_{\text{bbox, obj}}$	2560
$N_{\text{bbox, reg}}$	3669
Avg. S_{man}	0.2%
Avg. S_{obj}	0.7%
Avg. S_{reg}	4.5%

(c) Annotation metrics.

Figure 3. Statistics of MM-UAVBENCH.(a) Distribution of the 19 sub-tasks. (b) Proportions of the three input modalities. (c) Annotation metrics, where N_{video} and N_{img} denote the numbers of video clips and images; Avg. Res. denotes the average resolution; $N_{\text{bbox, man}}$, $N_{\text{bbox, obj}}$, and $N_{\text{bbox, reg}}$ denote the numbers of human, object, and region bounding boxes; Avg. S_{man}, Avg. S_{obj}, and Avg. S_{reg} represent their average area ratios.

uniformly downsample frames to 12 fps, which both simplifies manual annotation and aligns with mainstream practices in MLLM-based video processing.

3.2.2. Question-Answer Annotation

We adopt two approaches to construct MM-UAVBENCH: (1) direct human annotation, and (2) rule-based transformation from existing datasets. The detailed construction procedures for all 19 tasks are provided in the Appendix.

Human Annotation. For most tasks, annotators are provided with predefined task templates and annotate according to appropriate data sources (see Appendix for details). However, for perception-oriented tasks, relying solely on "templates + annotator judgment" is insufficient to control task difficulty. For example, in scene classification, MLLMs may already achieve high accuracy because similar scenes commonly appear in their pretraining corpora.

To address this, we first employ Qwen2.5-VL-72B to synthesize scene classification questions, deliberately increasing task difficulty by enriching the options with fine-grained details. Next, we use Qwen2.5-VL-7B and Qwen2.5-VL-72B answer these synthesized questions and select the cases where the two models disagree to human annotators. This way raises task difficulty from the data perspective. Furthermore, for tasks where options themselves are complex (e.g., the planning tasks where options may correspond to textual descriptions of distinct routes), we leverage MLLMs to expand the options, enhancing the plausibility of distractors.

Transfer from existing datasets. For datasets involving anomaly events and natural disasters, the original annotations usually contain rich semantic and structural information. We first apply rule-based methods to automatically synthesize multiple-choice questions for counting tasks and scene damage assessment. After expert verification, the questions are further refined by MLLMs to fit our task design.

3.2.3. Quality Control

The quality control of MM-UAVBENCH consists of two key aspects: (1) Annotation Accuracy Control, which ensures the consistency and reliability of annotations; and (2) Task Difficulty Control, which maintains a reasonable challenge level for both humans and models.

For annotation accuracy control, all tasks in MM-UAVBENCH originate from human annotations. Even for tasks adapted from existing datasets, we only retain samples that have been manually annotated or verified. Each sample is cross-checked by at least two professional researchers to ensure correctness and reduce annotation bias. Given the complexity of UAV scenes and task instructions, annotators may still disagree on certain samples. To mitigate this, domain experts developed detailed annotation guidelines grounded in scene semantics and the functional roles of UAVs. For factual tasks (i.e., tasks involving events that objectively occur in the video), we further standardize the answering viewpoint. For example, in orientation or road-related descriptions, the annotation protocol explicitly specifies whether the reference frame is the UAV itself or a ground agent. This is essential as prior work shows that MLLMs often produce inconsistent results across different viewpoints of the same question [27].

For task difficulty control, we adjust the challenge level through systematic distractor design. In factual tasks, distractors are selected from objects that co-occur with the target in keyframes or share similar appearances and states, typically occupying less than 10%—and in most cases under 1%—of the image area. In hypothetical tasks, large models are employed to assist human annotators, but the generated distractors often lack discriminative strength. To mitigate this, we strictly control the granularity of answers and distractors during annotation and review, ensuring they focus on clearly distinguishable factors such as direction, angle, or object choice (e.g., the distinction between Option A and B), thereby maintaining task discriminability and validity.

Table 2. **Experimental results on MM-UAVBENCH.** Dark Orange indicates the best result among all models and light Orange indicates the best result among open-source models. ‡: We conduct human evaluation on a randomly chosen 10% subset of the questions from each task.

Methods	Rank	Avg.	*Scene. Class.*	*Orient. Class.*	*Env. State*	*Urban OCR*	*CA. Count*	*RE. Count*	*Target Back.*	*Cross-Obj R.*	*Intent Anal.*	*Scene Attri.*	*Scene Damage*	*Scene Pred.*	*Event Trace*	*Event Under.*	*Event Pred.*	*Temporal Order*	*Swarm Plan*	*Ground Plan*	*Air-Ground Plan*
			Perception						Cognition										Planning		
Baseline																					
Random	-	25.38	25.00	25.00	25.00	25.00	20.00	20.00	26.48	28.15	25.00	32.55	25.00	25.00	24.73	25.00	24.40	25.00	28.07	25.00	22.91
Human‡	-	80.39	82.61	81.94	81.62	76.67	42.42	29.41	77.50	94.12	88.72	89.78	87.80	87.50	82.61	96.67	100.00	81.78	85.71	78.26	82.35
API-based																					
Gemini 2.5 Pro	1	54.59	74.90	37.89	73.78	82.19	23.94	24.86	51.69	48.26	50.00	84.12	44.57	57.00	73.25	73.09	68.02	51.14	25.68	44.19	48.56
Gemini 2.5 Flash	2	47.44	70.71	41.00	68.30	75.94	24.55	32.86	38.98	52.91	20.97	83.39	46.86	41.55	70.78	69.44	66.40	21.14	15.54	39.94	20.19
GPT-4o	3	44.92	57.32	38.43	68.01	62.19	22.12	18.86	14.41	25.58	50.00	83.39	31.71	40.10	67.08	64.78	63.16	33.71	28.38	39.09	45.19
Open-source																					
Qwen3-VL-8B	6	50.98	69.87	26.52	71.76	79.69	38.79	18.57	50.00	57.56	58.18	85.56	34.00	53.14	58.44	62.46	61.54	30.00	27.12	45.61	39.90
Qwen3-VL-32B	1	55.40	68.62	41.81	75.22	76.56	45.45	26.57	42.37	54.07	50.32	87.36	40.00	64.25	64.61	69.77	71.66	42.00	37.63	50.14	44.23
Qwen3-VL-235B-A22B	2	55.07	69.04	41.14	75.79	73.44	47.27	26.00	43.22	51.74	51.94	86.28	39.43	58.45	69.96	72.43	68.42	40.57	33.45	46.74	50.96
Qwen2.5-VL-7B	7	49.64	68.62	31.12	67.72	70.31	30.00	27.71	50.00	51.16	51.94	87.73	29.71	51.21	53.91	57.81	60.73	35.43	27.12	49.01	41.83
Qwen2.5-VL-32B	5	52.02	66.95	29.63	68.88	76.56	48.79	19.14	36.44	47.67	52.90	89.53	40.29	64.25	66.67	61.13	59.11	39.43	23.73	51.56	45.67
Qwen2.5-VL-72B	3	54.62	70.29	34.78	71.76	75.62	35.45	24.29	38.14	50.58	57.74	89.17	25.14	64.25	71.60	66.45	72.87	56.12	35.25	50.71	47.60
InternVL3.5-8B	9	47.13	58.16	27.06	72.62	68.12	36.36	17.71	34.75	48.84	36.77	88.09	34.00	49.28	54.73	56.26	60.73	34.57	40.68	39.66	37.02
InternVL3.5-38B	13	43.45	46.86	22.33	55.04	66.56	49.09	18.29	27.97	29.65	47.62	63.18	24.57	52.17	67.08	59.47	63.56	35.14	32.54	24.08	40.38
InternVL3-14B	10	46.86	66.53	18.40	72.05	73.75	32.42	16.29	41.53	54.07	33.87	86.28	42.00	56.52	58.44	47.84	54.25	35.71	29.49	36.26	34.62
InternVL3-78B	4	53.56	72.80	52.23	66.86	58.44	45.45	16.00	50.00	45.93	40.65	88.09	48.86	59.42	73.25	69.77	69.64	40.86	38.64	46.18	34.62
LLaVA-OneVision-7B	12	43.83	62.76	18.13	69.74	65.94	25.15	24.29	40.68	48.84	42.26	55.60	27.71	31.88	64.20	55.81	55.47	28.86	27.46	41.36	46.63
MiniCPM-V-4.5-8B	8	47.70	63.18	36.27	72.05	72.50	38.79	27.14	38.14	44.77	48.06	56.68	34.57	38.16	60.49	60.80	56.28	33.14	39.32	42.78	43.27
MiMo-VL-7B-RL	11	44.33	67.36	24.22	68.88	70.62	16.97	31.43	43.22	32.56	48.06	84.48	29.14	54.11	40.74	39.20	36.84	40.00	23.73	48.44	42.31

3.3. Diversity Statistics

The statistical overview of MM-UAVBENCH is presented in Figure 3. MM-UAVBENCH consists of 19 sub-tasks with a total of 5702 QA pairs. Among them, 82% are manually annotated, while the remaining 18% are converted from publicly available datasets that were originally human-annotated. The benchmark covers three input modalities: single images, key frames, and videos, their distribution is shown in Figure 3b. In total, the annotated data span 1549 video clips and 2873 images, with an average resolution of 1622×1033. The maximum resolution reaches 5472×3648. We overall annotate 7496 bounding boxes across three categories—regions, objects, and humans—whose average areas account for 4.5%, 0.7%, and 0.2% of the corresponding input frame, respectively. Taken together, these statistics demonstrate that MM-UAVBENCH poses a comprehensive and challenging evaluation for MLLMs across diverse tasks, and real-world complexities.

4. Experiment

4.1. Settings

Metrics. All questions in MM-UAVBENCH are designed in a multiple-choice format. We report accuracy as the primary evaluation metric. Each model is evaluated three times, and the average accuracy is taken as the final score for each task. For reproducibility, we use a greedy decoding configuration with top_p = 1.0, temperature = 0.0, and num_beams = 3.

Baselines. We select representative proprietary and open-source MLLMs as our baseline models. For the proprietary category, we include state-of-the-art models such as GPT-4o [12], Gemini 2.5 Pro and Gemini 2.5 Flash [3]. For the open-source category, we adopt Qwen3-VL and Qwen2.5-VL series [1], InternVL3.5 and InternVL3 series [33, 42], MiniCPM-V 4.5 [39, 40], LLaVA-OneVision-7B [16] and MiMo-VL-7B-RL [29]. More details of the evaluation are provided in Appendix.

4.2. Quantitative Results

The quantitative result is shown in Tab. 2. We rank the performance of all evaluated MLLMs and highlight the best scores in color. The main conclusions are summarized as follows:

- **Limited adaptability of current MLLMs to UAV scenarios.** While human evaluators achieve 80.4% average accuracy on our benchmark, existing MLLMs still struggle to adapt effectively to low-alititude UAV tasks. Notably, human performance achieve high scores on cognition and planning tasks, ranging from 78% to 100%, however except one task in cognition, the best scores of MLLMs

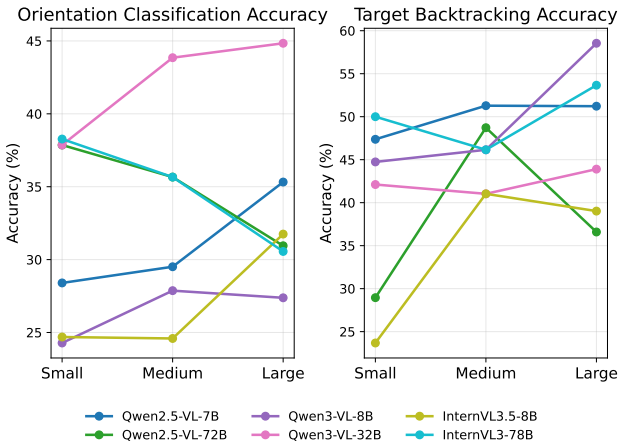


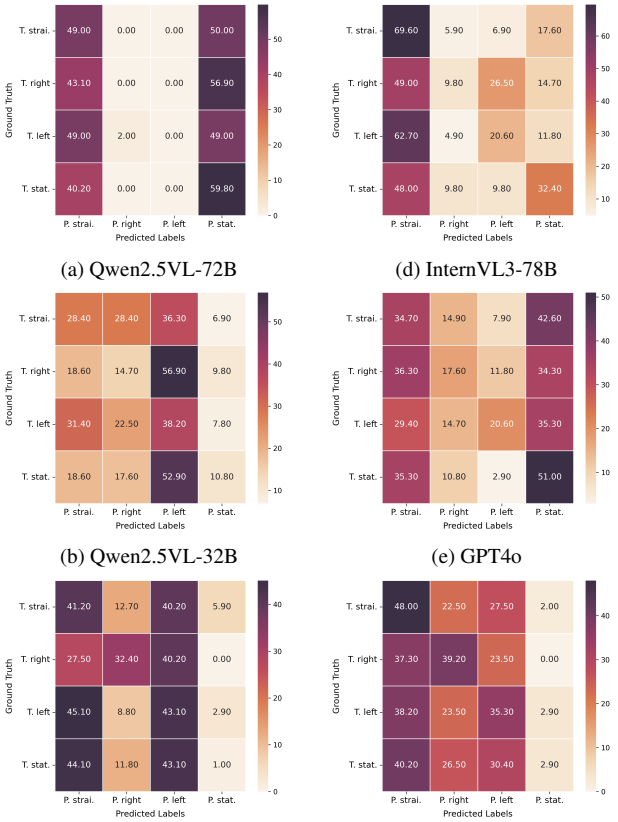
Figure 4. Accuracy comparison across small, medium, and large target sizes on Orient. Classification and Target Backtracking tasks.

ranging from 40% to 73%. Among all models, Gemini 2.5 Pro achieves the best overall performance, while Qwen3-VL-32B ranks highest among open-source models.

- **Proprietary and open-source MLLMs perform comparably.** Our results show that the performance gap between proprietary and open-source MLLMs is not significant. For example, the overall performance of Gemini 2.5 Flash and GPT-4o is around the median among open-source models. This suggests that the challenges posed by UAV scenarios are a general issue for current MLLMs, regardless of whether they are proprietary or open-source.
- **Model size influences performance.** We observe a clear trend that models with larger parameter scales tend to achieve higher accuracy, whereas smaller models generally perform worse. This finding highlights a potential trade-off between performance and deployability for MLLMs in low-altitude UAV scenarios.

4.3. Influence of Object Scale

We further analyze how object scale affects model performance. For the two tasks with annotated target bounding boxes—Orientation Classification and Target Backtracking—we group all questions into three subsets (Small, Medium, Large) based on the size of the referenced target. We then compute the accuracy for each subset, as summarized in Figure 4. Overall, accuracy tends to improve from small to large targets across models, indicating that current MLLMs struggle when the object occupies only a small portion of the field of view. This trend highlights object scale as a key factor shaping model performance in UAV scenarios.



(a) Qwen2.5VL-72B (d) InternVL3-78B

(b) Qwen2.5VL-32B (e) GPT4o

(c) Qwen2.5VL-7B (f) MiniCPM-V-4.5

Figure 5. Confusion matrices of predicted (P.) directions from MLLMs versus ground-truth (T.) on the Orient. Classification task.

4.4. Spatial Prediction Bias in MLLMs

We randomly sample 100 instances from each of the four orientation categories in the Orientation Classification task and compare the model predictions with the ground-truth labels. As shown in Fig. 5, the confusion matrices reveal strong model-dependent biases in orientation prediction. Qwen2.5-VL-72B almost never predicts left or right turns, whereas Qwen2.5-VL-32B and Qwen2.5-VL-7B tend to favor turning left and going straight, respectively. Other models exhibit similarly skewed behaviors. This highlights the difficulty current MLLMs face in extracting reliable motion cues from UAV perspectives.

4.5. Challenges in Multi-View Understanding

We further decompose the multi-view Intent Analysis and Prediction task into its single-view counterparts to assess whether current MLLMs can effectively leverage complementary cross-view information. Specifically, for each multi-view sample, we evaluate the model independently on each available view to obtain its single-view performance. As shown in Table 3, multi-view performance for most models

Table 3. Accuracy of different MLLMs under single-view (Aerial, Ground) and multi-view settings on the Intent Analysis and Prediction task. We additionally report the performance gap between multi-view and the best single view in each group (Δ).

Model	Aerial & Ground Views				Aerial Multi-Views			
	Aerial	Ground	Both	Δ	View1	View2	Both	Δ
GPT-4o	44.12	61.76	57.35	-4.41	45.29	50.36	49.09	-1.27
Qwen3-VL-235B	52.94	55.88	47.06	-8.82	54.35	55.07	52.54	-2.53
Qwen3-VL-32B	67.65	52.94	61.76	-5.89	51.09	48.91	48.91	-2.18
Qwen2.5-VL-7B	61.76	64.71	70.59	+5.88	48.91	48.91	49.64	+0.73
Qwen2.5-VL-72B	76.47	79.41	55.88	-23.53	35.14	35.51	57.97	+22.46
InternVL3.5-8B	44.12	38.24	38.24	-5.88	36.59	36.96	36.59	-0.37
InternVL3.5-38B	35.29	32.35	35.29	0.00	42.03	45.29	42.75	-2.54
InternVL3-78B	44.12	41.18	44.12	0.00	44.57	41.67	40.22	-4.35
MiniCPM-V-4.5	52.94	61.76	52.94	-8.82	54.35	51.81	51.45	-2.90
MiMo-VL-7B-RL	64.71	64.71	55.88	-8.83	47.83	49.28	47.10	-2.18

Table 4. Performance comparison between egocentric and exocentric planning, obtained by decomposing the original Air–Ground Collaborative Planning task (Mixed).

Model	Egocentric Plan.	Exocentric Plan.	Mixed
Qwen3-VL-8B	50.00	**59.18**	34.44
Qwen3-VL-32B	60.20	**65.31**	36.67
Qwen2.5-VL-72B	56.12	**58.16**	30.00
InternVL3.5-8B	46.94	**52.41**	30.00
InternVL3.5-38B	44.90	**47.96**	38.89
InternVL3-78B	**54.08**	51.02	28.89
MiMo-VL-7B-RL	54.08	**60.20**	26.67

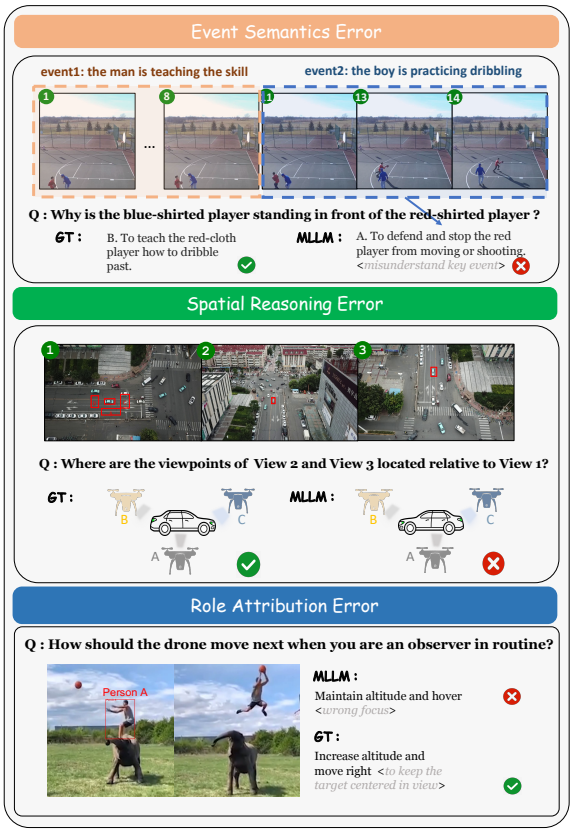


Figure 6. Qualitative failure analysis of MLLMs on MM-UAVBENCH.

does not surpass the best single-view result. The Δ column clearly indicates that, in both the Aerial–Ground and Aerial Multi-View settings, multi-view accuracy is frequently lower than that of the strongest single view—revealing a distinct "$1 + 1 < 2$" effect. Only a few models (e.g., Qwen2.5-VL-7B and Qwen2.5-VL-72B in the aerial multi-view case) achieve positive gains, while the majority show negative or marginal improvements. These findings demonstrate that current MLLMs lack effective view fusion and fail to combine complementary perspectives into stronger predictions.

4.6. Difficulty in Egocentric Planning

The Air–Ground Collaborative Planning task requires UAVs to simultaneously plan for ground agents and for their own self-motion. To better examine model behavior, we decompose this task into two subcomponents: exocentric planning, which focuses on predicting or planning for ground objects or other agents, and egocentric planning, which concerns the UAV's own goal-directed decision-making. As shown in Table 4, models consistently perform better on exocentric than on egocentric planning, indicating that they are more adept at interpreting external scene dynamics than reasoning about their own actions. Moreover, performance in the Mixed setting, where both self- and other-centric cues must be integrated, is substantially lower across all models, revealing significant difficulty in combining these two forms of reasoning. These results highlight a fundamental gap in current MLLMs that they struggle to ground predictions in their own embodiment and to generate coherent plans.

4.7. Other Error Analysis

We further perform qualitative analysis and identify three additional types of failures beyond the earlier quantitative findings, as illustrated in Figure 6:

- **Event Semantics Error.** MLLMs fail to correctly interpret the core semantics of an event—such as who is teaching, practicing, attacking, or defending. Misjudging these key actions leads to incorrect understanding of the event dynamics and flawed subsequent reasoning.
- **Spatial Reasoning Error.** MLLMs misinterpret the spatial correspondence between 2D schematic layouts and 3D real-world configurations, resulting in incorrect judgments about UAV viewpoints, relative positions, and coverage relationships across multiple views.
- **Role Attribution Error.** MLLMs incorrectly assign semantic roles to the entities involved in the scene—for

example, confusing the person who should be tracked or misidentifying who serves as the primary actor. Such role attribution mistakes lead to incorrect predictions and misguided UAV planning decisions.

5. Conclusions

In this work, we introduce MM-UAVBENCH, a comprehensive benchmark designed to evaluate the perception, cognition, and planning capabilities of multimodal large language models in low-altitude UAV scenarios. MM-UAVBENCH offers a diverse, high-fidelity, and domain-tailored testbed for assessing MLLM performance. Through extensive evaluations and detailed analyses, we show that while current MLLMs exhibit promising general capabilities, they struggle with UAV-specific challenges such as object-scale variation, spatial perception bias, multi-view understanding, and egocentric planning. These findings highlight a clear gap between generic multimodal intelligence and the requirements of realistic UAV operations. We hope that MM-UAVBENCH will inspire future research toward more capable, reliable, and UAV-oriented MLLMs for real-world deployment.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 6

[2] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8504–8510. IEEE, 2020. 1, 3

[3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[4] Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16985–16995, 2024. 3

[5] Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1

[6] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 1, 3

[7] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 1

[8] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22678–22690, 2024. 3

[9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 3

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1, 3

[11] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. 1

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[13] Woosuk Kwon, Zhuohan Li, Siyuan Lin, Xingyu Li, Conglong Shen, Eric Michael, Zhuang Zhang, Xiaoxia Yan, Kriss Wang, Ying Zhang, et al. vllm: Easy, fast, and cheap llm serving with pagedattention. *arXiv preprint arXiv:2309.06173*, 2023. 1

[14] Christos Kyrkou and Theocharis Theocharides. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In *CVPR workshops*, pages 517–525, 2019. 3

[15] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 1, 3

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[17] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021. 1, 3

[18] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024. 1, 3

[19] Bo Liu, Pengfei Qiao, Minhan Ma, Xuange Zhang, Yinan Tang, Peng Xu, Kun Liu, and Tongtong Yuan. Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models. *arXiv preprint arXiv:2505.12589*, 2025. 3

[20] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. Tasselnet: Counting maize tassels in the wild via local counts regression network. *Plant methods*, 13(1):79, 2017. 3

[21] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020. 3

[22] Mengjingcheng Mo, Xinyang Tong, Jiaxu Leng, Mingpi Tan, Jiankang Zheng, Yiran Liu, Haosheng Chen, Ji Gan, Weisheng Li, and Xinbo Gao. A2seek: Towards reasoning-centric benchmark for aerial anomaly understanding. *arXiv preprint arXiv:2505.21962*, 2025. 1

[23] Lichao Mou, Yuansheng Hua, Pu Jin, and Xiao Xiang Zhu. Era: A data set and deep learning benchmark for event recognition in aerial videos [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(4):125–133, 2020. 3

[24] Hemal Naik, Junran Yang, Dipin Das, Margaret C Crofoot, Akanksha Rathore, and Vivek H Sridhar. Bucktales: A multi-uav dataset for multi-object tracking and re-identification of wild antelopes. *Advances in Neural Information Processing Systems*, 37:81992–82009, 2024. 1

[25] Maryam Rahnemoonfar, Tashnim Chowdhury, and Robin Murphy. Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment. *Scientific data*, 10(1):913, 2023. 3

[26] Wen Shao, Rei Kawakami, Ryota Yoshihashi, Shaodi You, Hidemichi Kawase, and Takeshi Naemura. Cattle detection and counting in uav images based on convolutional neural networks. *International Journal of Remote Sensing*, 41(1): 31–52, 2020. 3

[27] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering, 2025. 5

[28] Zhichao Sun, Yepeng Liu, Huachao Zhu, Yuliang Gu, Yuda Zou, Zelong Liu, Gui-Song Xia, Bo Du, and Yongchao Xu. Refdrone: A challenging benchmark for referring expression comprehension in drone scenes. *arXiv preprint arXiv:2502.00392*, 2025. 1, 3

[29] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, et al. Mimo-vl technical report, 2025. 6

[30] Yonglin Tian, Fei Lin, Yiduo Li, Tengchao Zhang, Qiyao Zhang, Xuan Fu, Jun Huang, Xingyuan Dai, Yutong Wang, Chunwei Tian, et al. Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility. *Information Fusion*, 122:103158, 2025. 2

[31] Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? *arXiv preprint arXiv:2503.23771*, 2025. 1, 3

[32] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7907–7915, 2025. 1

[33] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6

[34] Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087*, 2024. 1, 3

[35] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2021. 1

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Clement Chaumond, Geraldine Delangue, Anthony Moi, Pierric Cistac, Timothée Rault, Roman Santus, Stanislas Max, et al. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. 2

[37] Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjin Wu, Weili Guan, and Xiang Deng. Uav-on: A benchmark for open-world object goal navigation with aerial agents. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13023–13029, 2025. 1, 3

[38] Fanglong Yao, Yuanchang Yue, Youzhi Liu, Xian Sun, and Kun Fu. Aeroverse: Uav-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models. *arXiv preprint arXiv:2408.15511*, 2024. 3

[39] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, et al. Efficient gpt-4v level multimodal large language model for deployment on edge devices. *Nature Communications*, 16 (1):5509, 2025. 1, 6

[40] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. 6

[41] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. *arXiv preprint arXiv:2503.06157*, 2025. 1, 3

[42] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1, 6

[43] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018. 1

[44] Pengfei Zhu, Jiayu Zheng, Dawei Du, Longyin Wen, Yiming Sun, and Qinghua Hu. Multi-drone-based single object tracking with agent sharing network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):4058–4070, 2020. 3

[45] Pengfei Zhu, Tao Peng, Dawei Du, Hongtao Yu, Libo Zhang, and Qinghua Hu. Graph regularized flow attention network for video animal counting from drones. *IEEE Transactions on Image Processing*, 30:5339–5351, 2021. 3

[46] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 1, 3

MM-UAVBENCH: How Well Do Multimodal Large Language Models See, Think, and Plan in Low-Altitude UAV Scenarios?

Supplementary Material

6. Appendix Outline

In the supplementary materials, we provide additional details, results, and visualizations to complement the main paper:

- **MM-UAVBENCH Details.** Including L3 sub-task definitions, task templates for annotation, and other annotation details.
- **Evaluation Details.** Including full experimental setup, evaluation prompts, and post-processing after evaluation.
- **Extended Results and Analysis.** Including results on L2 category, results with CoT, more analysis on spacial prediction bias.
- **Visualizations and Challenging Cases.** Including challenging examples from each MM-UAVBENCH task and the corresponding MLLM responses.

7. MM-UAVBENCH Details

7.1. Definition of Each Task

The hierarchical categories and definition of each task are shown in Table 5. This taxonomy provides a comprehensive coverage of UAV-related MLLM capabilities, ranging from basic perception and scene understanding to complex event reasoning and collaborative planning.

7.2. Annotation Details

MM-UAVBENCH consists of 82% manually annotated tasks, while the remaining 18% are automatically converted from publicly available dataset.

For manually annotated tasks, annotators are divided into three groups according to the L1 task categories. Each group is provided with a customized task template. After understanding the corresponding atomic task, annotators select suitable data instances from the collected datasets and start the annotation process. The prototypes of each task and their associated data sources are summarized in Table 6.

The tasks derived from publicly datasets are Class-agnostic Counting, Referring Expression Counting, and Scene Damage Assessment. For counting tasks, we convert raw count annotations into multiple-choice questions by generating distractors with controlled deviations determined by the difficulty setting (e.g., if the ground truth $c \in [20, 50]$, distractors are generated as $c \pm 0.15c$, $c \pm 0.3c$.). We then randomize the placement of the correct answer to avoid positional bias, ensuring that all options remain plausible while preserving fine-grained difficulty control. For Scene Damage Assessment, we first established a quantitative metric by aggregating and weighting semantic damage descriptions from original annotations to derive a numerical damage score for each image. This score is mapped to four distinct severity levels ('No Damage', 'Minor Damage', 'Major Damage', and 'Total Destruction'). The task is divided into two sub-challenges: single-image assessment, where the model predicts the damage level of an individual image, and comparative ranking, where the model ranks the severity of damage across a set of three images.

8. Evaluation Details.

We evaluate all models using VLMEvalKit[7]. Our benchmark includes three input modalities: single image, key frames, and videos. For video-based tasks, the original videos are sampled at 3.0 fps. The evaluation prompt is provided below:

Evaluation Prompt

You are an expert in the field of drones. Please answer the following questions based on your professional knowledge.

For images or key frames input:
You have been provided with several images and a multiple-choice question related to the image.

For video input:
You have been provided with {len(frames)} separate frames uniformly sampled from a video and a multiple-choice question related to the video. The frames are provided in chronological order of the video.

Your task is to carefully analyze the input data to answer the question, choosing from the options provided. Respond with only the letter of the correct option.

Question: {Question}
Options: {Options}

Please select the correct answer from the options above.

We further use 'exact_matching' policy to extract response from the generated outputs. For efficiency in evaluating baseline models, we utilized vLLM[13] to ac-

Table 5. Task taxonomy of MM-UAVBENCH, including hierarchical categories, definitions, and examples.

L1 Category	L2 Category	L3 Sub-task	Task Definition	Example
Perception	Classification	Scene Classification	Capture and categorize scenes from the entire image or selected areas.	Observe the image. What is the primary type of scene within the red box area?
		Orientation Classification	Identify the ongoing turning behavior of vehicles/people.	Based on visible cues, what is the immediate motion direction of the white SUV within the red box area relative to its own orientation?
		Environment State Classification	Identify the lighting and weather conditions of the scene in the image.	What are the lighting and weather conditions in this image?
	OCR	Urban OCR	Recognize the text within the specified bounding box in the image.	What is the exact text on the warning sign at the extreme angle?
	Counting	Class-agnostic Counting	Count the objects of the specified category in the image.	Please count the number of the objects in the image that belong to the category: sheep
		Referring Expression Counting	Count the objects that meet the specific descriptions in the image.	Please count the number of the objects or people that match the description: The white vehicles waiting at the traffic light.
Cognition	Object-Level Reasoning	Target Backtracking	Trace back the spatial positions and behaviors of the target in the past spatiotemporal sequence.	What did the car do before reaching this point?
		Cross-Object Reasoning	Analyze the behavioral relationships or spatial connections among multiple subjects at the current moment.	Can another car be parked between these two cars?
		Intent Analysis and Prediction	Predict the target's future spatial positions and behaviors based on its current spatial position and behavior.	How will the car in the bounding box travel at the T-junction based on ground and aerial perspectives?
	Scene-Level Reasoning	Scene Attribute Understanding	Analyze the attributes or functions of the entire image or the selected region.	How can you describe this scene, Modern or retro?
		Scene Damage Assessment	Analyze the damage degree of the scene/Compare the damage degrees of multiple scenes.	How to rank the severity of building damage across the three scenes (from mild to severe)?
		Scene Analysis and Prediction	Predict the future change trends based on the changes occurring within the scene over a period of time.	Predict the most likely changes in traffic flow within the scene.
	Event-Level Reasoning	Event Tracing	Analyze the causes of the event's occurrence.	What led to the black cars closely follow white cars shown in the video?
		Event Understanding	Understand the events that are occurring in the video.	What specific harvesting operation is depicted in the aerial view of the cornfield?
		Event Prediction	Predict the future development trend of the events in the video.	How will the conflict situation shown in the video develop?
		Temporal Ordering	Analyze the chronological order of multiple key frames in a video.	In which order should these images be arranged to match the actual progression of the event?
Planning	UAV-to-UAV Planning	Swarm Collaborative Planning	Based on the information provided by multiple drones from different perspectives, select the optimal drone task allocation strategy.	There are three drones providing main, secondary, and third-perspective views. The drone offering the third perspective has withdrawn due to low battery. To maintain the original tracking plan, determine whether a new drone should be deployed and where it should be added.
	UAV-to-Ground Planning	Ground-Target Planning	The ground target needs to complete a specific task; provide a reasonable action plan or route.	If rescuers need to rescue the injured at ABCDE's point after the earthquake and take the wounded to the Region $\langle A \rangle$ for evacuation, please plan the most suitable rescue route.
		Air-Ground Collaborative Planning	Provide a reasonable action plan or route for the ground target and drones to jointly complete a specific task.	A religious activity is in progress. Based on the information boxed in the picture, to ensure that Vehicle B can move forward smoothly, what actions should Vehicle A and the drone take?

celerate inference for models based on the Qwen architecture, while other models were run using the standard Hugging Face transformers library[36]. To ensure reproducibility, we strictly set the generation configuration with $temperature = 0.0$ and $top\_p = 1.0$. Furthermore, where supported by the model, we optionally employed $num\_beams = 3$ for generation.

9. Extended Results and Analysis.

Here we provide more evaluation results and analysis on MM-UAVBENCH, including L2 category results, results of models with Chain-of-Thought(CoT), and more analysis on spacial prediction bias.

Table 6. Task templates and annotation data sources used for each L3 sub-task.

L3 Sub-task	Task Template	Data Source
Scene Classification	Given an image or a selected region within the image, what category does the corresponding scene belong to?	Visdrone-DET[46]
Orientation Classification	Given an image and the selected object, what is its current turning intention based on its own movement direction?	Visdrone-DET[46]
Environment State Classification	Given an image, what are the lighting and climate conditions of it?	Visdrone-DET[46], Visdrone-VID[46], ERA[23], AIDER[14], MDOT[44]
Urban OCR	Given an image, recognize the text within the selected region.	Visdrone-DET[46]
Class-agnostic Counting	Count the number of the specified type of object in a given image.	Animaldrone[45], Cattle-det[26], MTC-plant[20]
Referring Expression Counting	Count the number of objects that match the referring expression in a given image.	Refdrone[28], Rec8k[4], Visdrone-DET[46]
Target Backtracking	Given multiple key frames in a video, what are the [spatial position] or [behavior] of the target object or person in the [past spatiotemporal sequence]?	Visdrone-VID[46], Visdrone-SOT[46]
Cross-Object Reasoning	Given an image and multiple target objects or people, what are the [behavioral relationship] or [spatial relationship] between them?	Visdrone-DET[46]
Intent Analysis and Prediction	Given multiple frames in a video and a target object or person, based on its [spatial position and behavior] [from the past to the present], what will its [future] [spatial position and behavior] be?	Mavrec[8], MDOT[44]
Scene Attribute Understanding	Given an image or a selected region within the image, does it conform to a certain [description]? / what is its [function]?	Visdrone-DET[46]
Scene Damage Assessment	Given an image, how severe is the [disaster level] of the scene it shows? / Given multiple images, sort them by the severity of their disaster levels.	RescueNet[25]
Scene Analysis and Prediction	Given multiple frames in a video, based on the changes the scene has undergone [from the past to the present]. What its [future change trend] be?	UAVid[21], AU-AIR[2]
Event Tracing	Given multiple frames in a video, analyze the [causes of the event] happened in the [past spatiotemporal sequence]?	ERA[23]
Event Understanding	Given multiple frames in a video, understand the [event] that is [currently] happening.	ERA[23]
Event Prediction	Given multiple frames in a video, based on the [event] happened [from the past to the present], predict the [future evolution] of the event.	ERA[23]
Temporal Ordering	Given multiple [shuffled frames] of a video, based on the [stages of events] in different frames, what is the [correct chronological order] of the frames	ERA[23]
Swarm Collaborative Planning	Given multiple images from [multiple drone perspectives], select the perspective with the most comprehensive information as the main perspective and mark [multiple candidate regions in the main perspective]. Under a specific requirement, which region should [be prioritized for allocation]? / If the drone corresponding to the perspective of a certain image is [damaged], which region needs to have [a new drone added]?	MDOT[44]
Ground-Target Planning	Given an image of a [disaster-affected scene] and [multiple marked rescue points], what is the [most suitable rescue route] and plan considering both rescue [priority] and rescue [time]?	AIDER[14]
Air-Ground Collaborative Planning	Given multiple frames in a video, assuming [a task] needs to be performed, what actions should the [ground targets] and [UAVs] take respectively based on the current state?	ERA[23]

Table 7. Averaged zero-shot evaluation results on MM-UAVBENCH on L2 Category.

Methods	Classification	OCR	Counting	Object	Scene	Event	UAV-to-UAV	UAV-to-Ground
	Perception			Cognition			Planning	
API-based								
Gemini 2.5 Pro	62.19	82.19	24.40	50.05	61.89	66.62	25.68	46.38
Gemini 2.5 Flash	60.00	75.94	28.71	37.62	57.27	56.44	15.54	30.07
GPT-4o	54.59	62.19	20.49	30.00	51.74	57.18	28.38	42.14
Open-source								
Qwen3-VL-8B	56.05	79.69	28.68	55.25	57.57	53.11	27.12	42.76
Qwen3-VL-32B	61.88	76.56	36.01	48.90	63.87	61.96	37.63	47.19
Qwen3-VL-235B-A22B	61.99	73.44	36.64	48.97	58.09	62.84	33.45	48.85
Qwen2.5-VL-7B	55.82	70.31	28.86	49.34	56.22	51.97	27.12	45.42
Qwen2.5-VL-32B	55.15	76.56	36.01	45.69	64.69	56.59	23.73	48.62
Qwen2.5-VL-72B	58.94	75.62	29.87	48.82	59.85	66.21	35.25	49.16
InternVL3.5-8B	52.61	68.12	27.04	40.12	57.02	51.34	40.68	38.34
InternVL3.5-38B	41.38	66.56	33.69	33.43	46.61	56.31	32.54	32.23
InternVL3-14B	52.33	73.75	24.36	43.12	61.53	51.58	29.49	35.44
InternVL3-78B	63.96	58.44	30.73	45.59	65.43	63.38	38.64	40.40
LLaVA-OneVision-7B	50.21	65.94	24.72	37.26	38.40	51.09	27.46	44.00
MiniCPM-V-4.5-8B	57.17	72.50	32.96	43.66	43.14	52.56	39.32	43.03
MiMo-VL-7B-RL	53.49	70.62	24.20	41.28	55.89	39.19	23.73	45.38

9.1. Results on L2 Category

Table 7 presents the zero-shot evaluation results of 16 general MLLMs across the L2 categories, which directly reveals specific capability deficiencies required for UAV scenarios. We can summarize limitations of MLLMs across three core dimensions:

- **Weakness in fine-grained quantitative perception.** In the Perception dimension, performance on Counting ($\sim 20-36\%$) is significantly lower than Classification and OCR ($\sim 50-80\%$). This disparity underscores a severe bottleneck where MLLMs struggle with fine-grained enumeration and density estimation in aerial imagery.
- **Object-level reasoning is more challenging.** The Cognition dimension reveals that object-level reasoning is notably weaker than both scene-level and event-level reasoning. This decline is possibly correlated with the small target scale and limited context. Fig. 4 also shows that Target Backtracking (an object-level reasoning task) improves with increasing target size.
- **Handling multi-view images increases the difficulty of planning.** MLLMs perform substantially worse on UAV-to-UAV collaborative tasks (multi-view input) than on UAV-to-Ground tasks (single-view input) in the Planning dimension. This strongly indicates MLLMs' fundamental limitation in processing and integrating information from disparate multi-view inputs, which is critical for complex swarm planning.

9.2. Evaluation with CoT

The performance gap between models with and without Chain-of-Thought (CoT) prompting is shown in Table 8. Overall, employing CoT significantly improves the average performance of both models (Qwen3-VL-8B:$\Delta + 2.55$; MiMo-VL-7B-RL:$\Delta + 5.55$). Across individual tasks, the gains are highly variable. CoT largely improves the performance in the perceptional tasks like Orientation Classification, Class-agnostic Counting, and most Event-Level tasks in cognition. Conversely, CoT does not perform well in Object-Level Reasoning tasks (e.g., Qwen3-VL-8B: $\Delta - 14.54$), suggesting that the explicit intermediate steps may introduce errors when the initial perception or localization of the target is inherently difficult.

9.3. Analysis on Spacial Prediction Bias

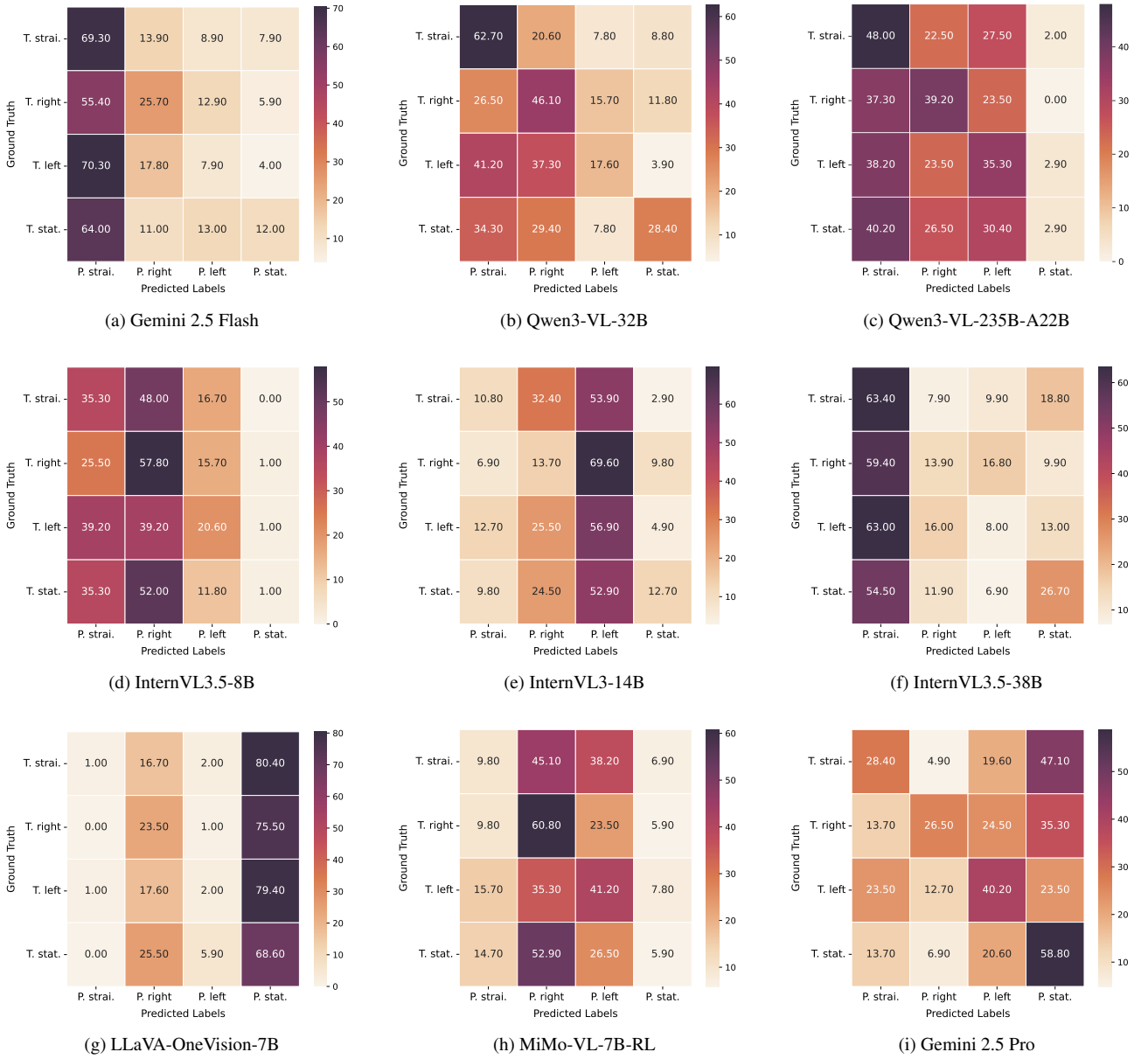
As shown in Fig. 7, we present the confusion matrices of other baseline models for the Orientation Classification task. Both API-based and open-source MLLMs exhibit spatial prediction biases. Specifically, models demonstrate a strong conservative bias, frequently misclassifying turning motions ('T. right', 'T. left') as the 'P. stat.' (predicted stationary) or 'T. strai.' (predicted straight) categories. This indicates that MLLMs struggle to accurately resolve subtle directional cues in aerial imagery, leading to confusion between rotational and static/forward movement intentions.

10. Visualizations and Challenging Cases

In this section, we present additional examples from MM-UAVBENCH along with the responses of baseline models. Representative cases for the Perception dimension are shown in Fig. 8 and Fig. 9, for the Cognition dimension in Fig. 10, Fig. 11, Fig. 12, Fig. 13 and Fig. 14, for the Planning dimension in Fig. 15 and Fig. 16.

Table 8. **Evaluation with CoT on L3 sub-tasks.** "Qwen3-VL-8B-Thinking" and "MiMo-VL-7B-RL-Thinking" denote the original models augmented with CoT. Δ represents the performance difference (CoT-augmented minus original) for each sub-task.

Methods	Avg.	Scene. Class.	Orient. Class.	Env. State	Urban OCR	CA. Count	RE. Count	Target Back.	Cross-Obj R.	Intent Anal.	Scene Attri.	Scene Damage	Scene Pred.	Event Trace	Event Under.	Event Pred.	Temporal Order	Swarm Plan	Ground Plan	Air-Ground Plan
		Perception						Cognition										Planning		
Qwen3-VL-8B	50.98	69.87	26.52	71.76	79.69	38.79	18.57	50.00	57.56	58.18	85.56	34.00	53.14	58.44	62.46	61.54	30.00	33.11	45.61	39.90
Qwen3-VL-8B-Thinking	53.54	69.46	45.20	71.47	72.50	47.88	24.56	40.17	43.02	52.68	84.06	41.33	66.67	63.79	65.78	65.59	39.47	36.21	46.31	41.06
Δ	+2.55	-0.41	+18.68	-0.29	-7.19	+9.09	+5.99	-9.83	-14.54	-5.50	-1.50	+7.33	+13.53	+5.35	+3.32	+4.05	+9.47	+3.10	+0.70	+1.16
Mimo-VL-7B-RL	44.67	67.36	24.22	68.88	70.63	16.97	31.43	43.22	32.56	48.06	84.48	29.14	54.11	40.74	39.20	36.84	40.00	30.08	48.44	42.31
Mimo-VL-7B-RL-Thinking	50.22	67.36	35.05	69.74	70.31	25.15	32.27	31.62	49.42	47.05	86.28	35.82	53.47	64.73	63.33	54.29	36.98	41.02	45.58	44.66
Δ	+5.55	-0.00	+10.83	+0.86	-0.32	+8.18	+0.84	-11.60	+16.86	-1.01	+1.80	+6.68	-0.64	+23.99	+24.13	+17.45	-3.02	+10.94	-2.86	+2.35



(a) Gemini 2.5 Flash (b) Qwen3-VL-32B (c) Qwen3-VL-235B-A22B

(d) InternVL3.5-8B (e) InternVL3-14B (f) InternVL3.5-38B

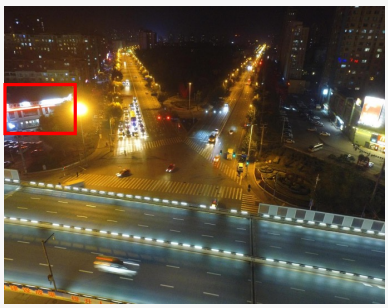
(g) LLaVA-OneVision-7B (h) MiMo-VL-7B-RL (i) Gemini 2.5 Pro

Figure 7. More confusion matrices of other baseline models on the Orientation Classification task.

Scene Classification

Q:What is the primary type of scene within the red box area?

A: Shopping mall
B: Hotel
C: Hospital
D: Bank



Gemini 2.5 Flash:	D ✓	Qwen3-VL-32B:	D ✓
GPT-4o:	A ✗	InternVL3.5-38B:	D ✓
Qwen3-VL-235B-A22B:	A ✗	InternVL3-78B:	D ✓
Gemini 2.5 Pro:	D ✓	LLaVA-OneVision-7B:	A ✗
Qwen2.5-VL-72B:	A ✗	MiniCPM-V-4.5_8B:	A ✗

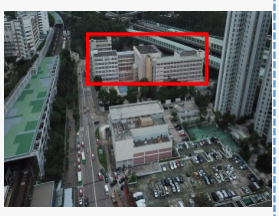
Q:What is the primary type of scene within the red box area?

A: Pond B: River
C: Sea D: Lake

Gemini 2.5 Flash:	D ✗
GPT-4o:	A ✗
Qwen3-VL-235B-A22B:	B ✓

Q:What is the primary type of scene within the red box area?

A: School
B: Residential buildings
C: Hospital
D: Department store building

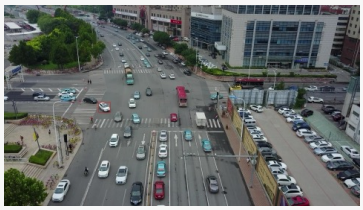


Gemini 2.5 Flash:	A ✓
GPT-4o:	B ✗
Qwen3-VL-235B-A22B:	B ✗

Orientation Classification

Q:Based on visible cues, what is the immediate motion direction of the white SUV within the red box area relative to its own orientation?

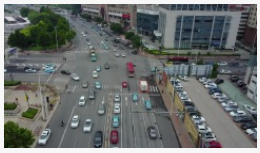
A: Moving straight
B: Turning right
C: Turning left
D: Stationary (no movement)



Gemini 2.5 Flash:	A ✗	Qwen3-VL-32B:	C ✗
GPT-4o:	C ✗	InternVL3.5-38B:	C ✗
Qwen3-VL-235B-A22B:	C ✗	InternVL3-78B:	D ✗
Gemini 2.5 Pro:	C ✗	LLaVA-OneVision-7B:	C ✗
Qwen2.5-VL-72B:	A ✗	MiniCPM-V-4.5_8B:	C ✗

Q:Based on visible cues, what is the immediate motion direction of the white car within the red box area relative to its own orientation?

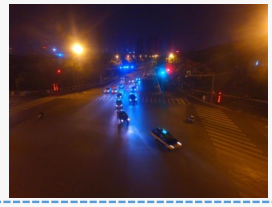
A: Moving straight B: Turning left
C: Turning right D: Stationary



Gemini 2.5 Flash:	A ✓
GPT-4o:	D ✗
Qwen3-VL-235B-A22B:	B ✗

Q:Based on visible cues, what is the immediate motion direction of the white car within the red box area relative to its own orientation?

A: Moving straight
B: Turning left
C: Turning right
D: Stationary (no movement)



Gemini 2.5 Flash:	A ✗
GPT-4o:	B ✗
Qwen3-VL-235B-A22B:	A ✗

Figure 8. Additional Examples of Sub-Tasks (Part 1).

Environment State Classification

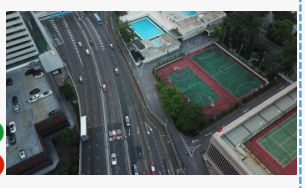
Q:What are the lighting and weather conditions in this image?

A: daytime and cloudy
B: nighttime and foggy
C: daytime and snowy
D: daytime and sunny

Gemini 2.5 Flash:	A ✗	Qwen3-VL-32B:	A ✗
GPT-4o:	A ✗	InternVL3.5-38B:	A ✗
Qwen3-VL-235B-A22B:	A ✗	InternVL3-78B:	A ✗
Gemini 2.5 Pro:	A ✗	LLaVA-OneVision-7B:	A ✗
Qwen2.5-VL-72B:	A ✗	MiniCPM-V-4.5_8B:	A ✗

Q:What are the lighting and weather conditions in this image?

A: dusk and clear
B: dusk and foggy
C: daytime and clear
D: daytime and cloudy



Gemini 2.5 Flash:	D ✓
GPT-4o:	C ✗
Qwen3-VL-235B-A22B:	C ✗

Q:What are the lighting and weather conditions in this image?

A: daytime and rainy
B: dusk and cloudy
C: nighttime and clear
D: dusk and clear

Gemini 2.5 Flash:	D ✗
GPT-4o:	D ✗
Qwen3-VL-235B-A22B:	D ✗

Urban OCR

Q:What is the exact text on the billboard?

A: 中联农仙
B: 中联农业
C: 中辰农业
D: 中联农先

Gemini 2.5 Flash:	B ✗	Qwen3-VL-32B:	D ✗
GPT-4o:	A ✗	InternVL3.5-38B:	A ✗
Qwen3-VL-235B-A22B:	A ✗	InternVL3-78B:	C ✓
Gemini 2.5 Pro:	C ✗	LLaVA-OneVision-7B:	D ✗
Qwen2.5-VL-72B:	A ✓	MiniCPM-V-4.5_8B:	A ✗

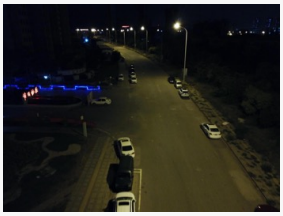
Q:How many seconds are left on the red straight-going light in the image?

A: 30 seconds
B: 39 seconds
C: 34 seconds
D: 38 seconds

Gemini 2.5 Flash:	C ✗
GPT-4o:	C ✗
Qwen3-VL-235B-A22B:	C ✗

Q:What is the exact text on the illuminated signboard?

A: 華國佳苑
B: 节日快乐
C: 華国住苑
D: 華国佳院



Gemini 2.5 Flash:	A ✗
GPT-4o:	A ✗
Qwen3-VL-235B-A22B:	A ✗

Figure 9. Additional Examples of Sub-Tasks (Part 2).

Target Backtracking

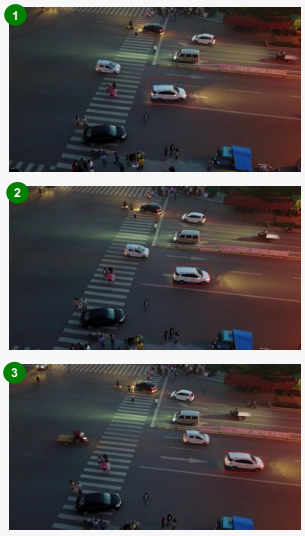
Q:Was the car within the red box area in motion before this moment?

A: Yes
B: No
C: Insufficient visual cues
D: Unrelated to the visual cues

Gemini 2.5 Flash:	C	✗
GPT-4o:	C	✗
Qwen3-VL-235B-A22B:	C	✗
Gemini 2.5 Pro:	A	✓
Qwen2.5-VL-72B:	C	✗
Qwen3-VL-32B:	C	✗
InternVL3.5-38B:	A	✓
InternVL3-78B:	A	✓
LLaVA-OneVision-7B:	B	✗
MiniCPM-V-4.5_8B:	C	✗

Q:Did the vehicle in the box turn at a right angle from the left intersection to the current position?

A: Yes
B: No
C: Insufficient visual cues
D: Unrelated to the visual cues



Gemini 2.5 Flash:	/	✗
GPT-4o:	B	✗
Qwen3-VL-235B-A22B:	D	✗
Gemini 2.5 Pro:	D	✓
Qwen2.5-VL-72B:	B	✗
Qwen3-VL-32B:	B	✗
InternVL3.5-38B:	B	✗
InternVL3-78B:	B	✓
LLaVA-OneVision-7B:	D	✗
MiniCPM-V-4.5_8B:	C	✗

Cross-Object Reasoning

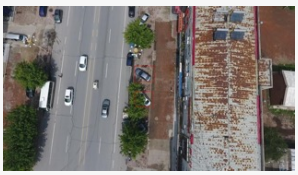
Q:What are the two people within the red box area doing in the picture?

A: walk side by side
B: pursue
C: roll about
D: jump

Gemini 2.5 Flash:	A	✗	Qwen3-VL-32B:	A	✗
GPT-4o:	B	✓	InternVL3.5-38B:	B	✓
Qwen3-VL-235B-A22B:	A	✗	InternVL3-78B:	A	✗
Gemini 2.5 Pro:	A	✗	LLaVA-OneVision-7B:	C	✗
Qwen2.5-VL-72B:	A	✗	MiniCPM-V-4.5_8B:	A	✗

Q:Can a car park in the gap between these two cars?

A: Yes
B: No
C: Insufficient visual cues
D: Unrelated to the visual cues



Gemini 2.5 Flash:	/	✗	Qwen3-VL-32B:	B	✗
GPT-4o:	B	✗	InternVL3.5-38B:	A	✓
Qwen3-VL-235B-A22B:	C	✗	InternVL3-78B:	B	✗
Gemini 2.5 Pro:	B	✗	LLaVA-OneVision-7B:	B	✗
Qwen2.5-VL-72B:	B	✗	MiniCPM-V-4.5_8B:	B	✗

Figure 10. Additional Examples of Sub-Tasks (Part 3).

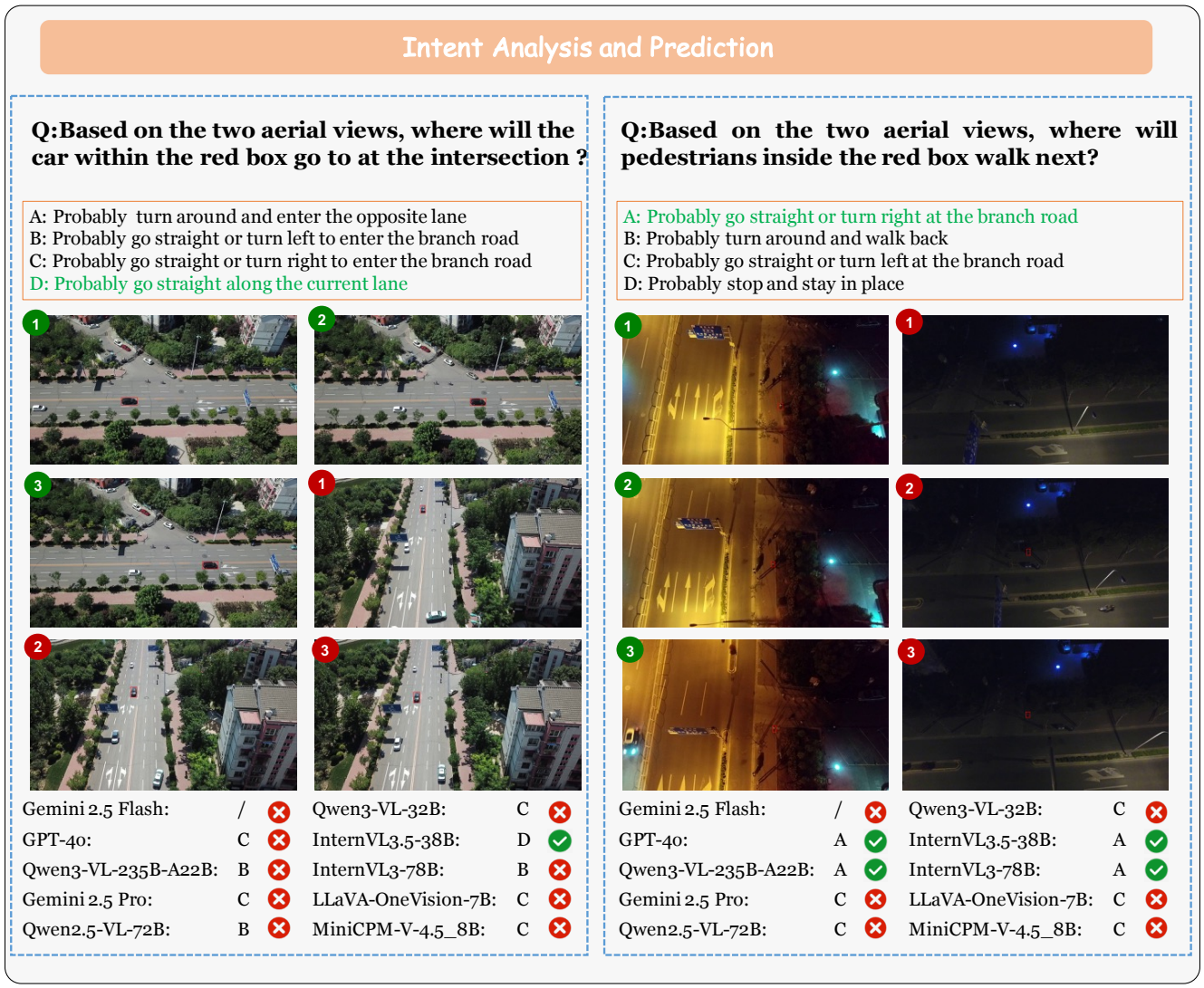
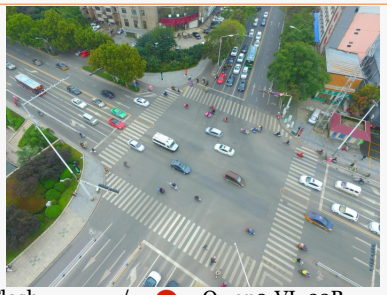


Figure 11. Additional Examples of Sub-Tasks (Part 4).

Scene Attribute Understanding

Q:What can vehicles do in the scene shown in the image from the camera's perspective?

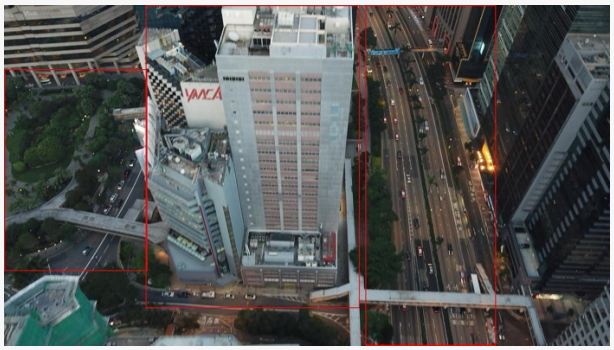
A: Vehicles on the inside lane of a longitudinal road running from bottom-to-top approach may turn left onto a transverse road running from right-to-left horizontal approach.
B: Vehicles on a transverse road running from right-to-left horizontal approach may turn right onto a longitudinal road running from top-to-bottom approach.
C: Vehicles on the inner lane of the longitudinal road running from top-to-bottom approach may turn right onto a transverse road running from right-to-left horizontal approach.
D: Vehicles on a transverse road running from left-to-right horizontal approach may remain in the intersection area.



Gemini 2.5 Flash:	/	✗	Qwen3-VL-32B:	C	✗
GPT-4o:	A	✓	InternVL3.5-38B:	C	✗
Qwen3-VL-235B-A22B:	A	✓	InternVL3-78B:	C	✗
Gemini 2.5 Pro:	C	✗	LLaVA-OneVision-7B:	D	✗
Qwen2.5-VL-72B:	C	✗	MiniCPM-V-4.5_8B:	C	✗

Q:What is the function of the scene shown in the middle of the image?

A: Used for residing and living.
B: Used for conducting commercial office work.
C: Used for conducting education and training.
D: Used for holding art exhibitions.



Gemini 2.5 Flash:	/	✗	Qwen3-VL-32B:	B	✗
GPT-4o:	B	✗	InternVL3.5-38B:	A	✗
Qwen3-VL-235B-A22B:	C	✓	InternVL3-78B:	B	✓
Gemini 2.5 Pro:	B	✗	LLaVA-OneVision-7B:	B	✓
Qwen2.5-VL-72B:	B	✓	MiniCPM-V-4.5_8B:	B	✓

Figure 12. Additional Examples of Sub-Tasks (Part 5).

Scene Damage Assessment

Q:For a scene-level assessment, choose the damage classification that fits the buildings as a whole in this image. (Base your answer on the dominant condition of the buildings; disregard unassigned wreckage or background clutter.)

A: No Damage
B: Minor Damage
C: Major Damage
D: Total Destruction

Gemini 2.5 Flash:	C ✓	Qwen3-VL-32B:	D ✗
GPT-4o:	D ✗	InternVL3.5-38B:	C ✓
Qwen3-VL-235B-A22B:	D ✗	InternVL3-78B:	C ✓
Gemini 2.5 Pro:	D ✗	LLaVA-OneVision-7B:	D ✗
Qwen2.5-VL-72B:	B ✗	MiniCPM-V-4.5_8B:	C ✓

Q:Arrange the scenes by increasing building damage severity (lowest → highest).

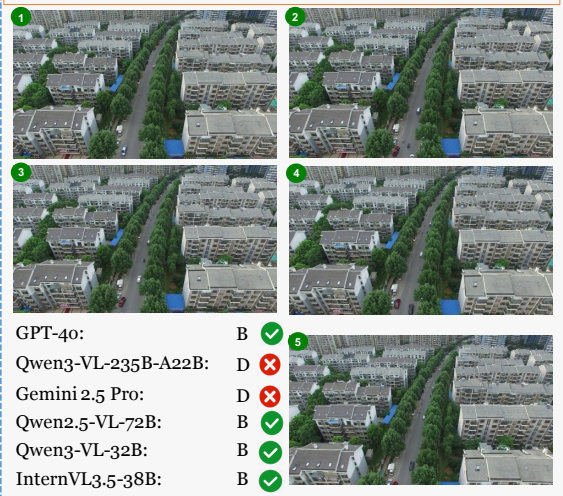
A: Scene1-Scene3-Scene2
B: Scene2-Scene3-Scene1
C: Scene3-Scene1-Scene2
D: Scene3-Scene2-Scene1

Gemini 2.5 Flash:	C ✗	Qwen3-VL-32B:	C ✗
GPT-4o:	A ✗	InternVL3.5-38B:	D ✓
Qwen3-VL-235B-A22B:	A ✗	InternVL3-78B:	C ✗
Gemini 2.5 Pro:	D ✓	LLaVA-OneVision-7B:	D ✓
Qwen2.5-VL-72B:	A ✗	MiniCPM-V-4.5_8B:	C ✗

Scene Analysis and Prediction

Q:Predict the possible changes that may occur in the future on both lanes.

A: The traffic volume in the left lane will increase significantly.
B: The traffic volume in both lanes will not show any significant changes.
C: The traffic flow in the right lane will decrease.
D: The traffic volume in both lanes will experience a significant increase.



GPT-4o:	B ✓	InternVL3-78B:	D ✗
Qwen3-VL-235B-A22B:	D ✗	LLaVA-OneVision-7B:	B ✓
Gemini 2.5 Pro:	D ✗	MiniCPM-V-4.5_8B:	D ✗
Qwen2.5-VL-72B:	B ✓		
Qwen3-VL-32B:	B ✓		
InternVL3.5-38B:	B ✓		

Q:Predict possible changes that may occur in the scene in the future.

A: The traffic flow will remain stable.
B: The traffic flow will gradually decrease, making this roundabout intersection more spacious.
C: The traffic flow at this roundabout intersection will increase and decrease from time to time.
D: The traffic flow will gradually increase, which may lead to frequent congestion at the roundabout intersections.

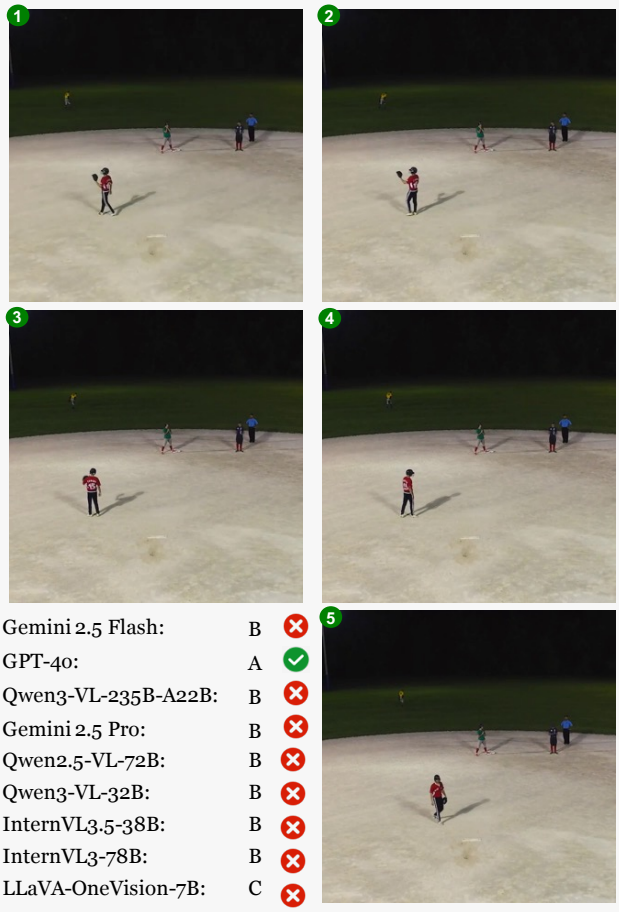
GPT-4o:	C ✗	InternVL3-78B:	B ✗
Qwen3-VL-235B-A22B:	C ✗	LLaVA-OneVision-7B:	D ✓
Gemini 2.5 Pro:	C ✗	MiniCPM-V-4.5_8B:	D ✓
Qwen2.5-VL-72B:	C ✗		
Qwen3-VL-32B:	D ✓		
InternVL3.5-38B:	C ✗		

Figure 13. Additional Examples of Sub-Tasks (Part 6).

Event Tracing

Q:What is the reason for the player in red moving left in the video?

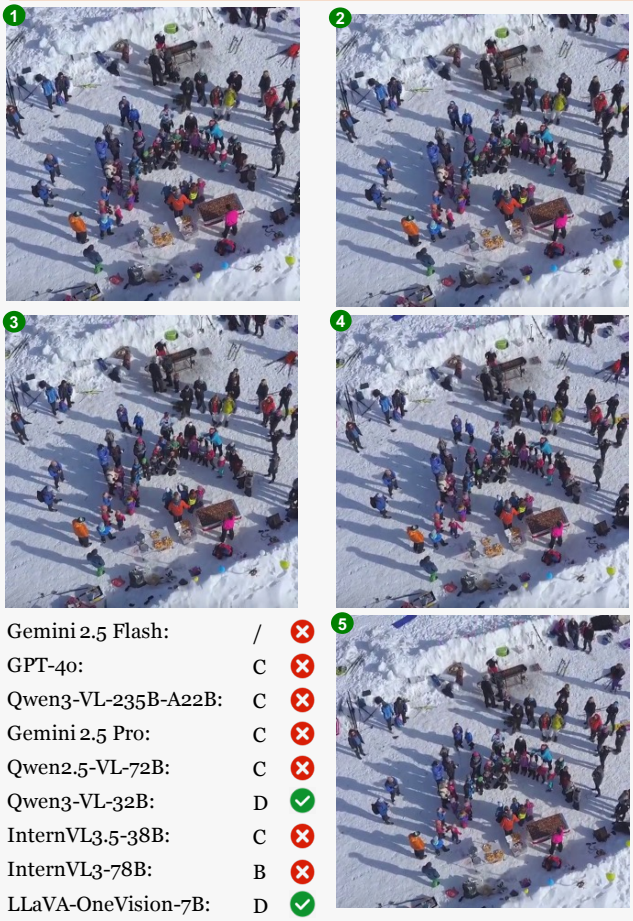
A: Prepare to receive the ball.
B: Adjust the position.
C: Prepare to pitch.
D: Dodge the incoming ball.



Gemini 2.5 Flash:	B	✗
GPT-4o:	A	✓
Qwen3-VL-235B-A22B:	B	✗
Gemini 2.5 Pro:	B	✗
Qwen2.5-VL-72B:	B	✗
Qwen3-VL-32B:	B	✗
InternVL3.5-38B:	B	✗
InternVL3-78B:	B	✗
LLaVA-OneVision-7B:	C	✗
MiniCPM-V-4.5_8B:	A	✓

Q:What was the reason for those people in the video to form a circle?

A: Because they are being filmed for a winter - themed documentary.
B: In order to conduct a teaching activity on winter survival skills.
C: To celebrate a child's birthday, a circle birthday event is held in the snow.
D: People are gathered in a circle to watch the performance.



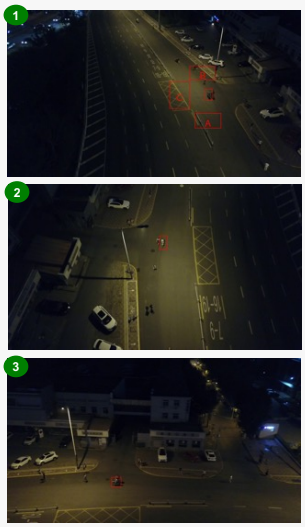
Gemini 2.5 Flash:	/	✗
GPT-4o:	C	✗
Qwen3-VL-235B-A22B:	C	✗
Gemini 2.5 Pro:	C	✗
Qwen2.5-VL-72B:	C	✗
Qwen3-VL-32B:	D	✓
InternVL3.5-38B:	C	✗
InternVL3-78B:	B	✗
LLaVA-OneVision-7B:	D	✓
MiniCPM-V-4.5_8B:	D	✓

Figure 14. Additional Examples of Sub-Tasks (Part 7).

Swarm Collaborative Planning

Q:At present, there are 3 drones, and the shooting angle is as shown in the figure, with the UAV in View 1 as the main command perspective, the drones in View 2 and View 3 provide second and third perspectives. Now you need to see the road ahead of the person you are tracking while tracking him, so in which area should you add a drone?

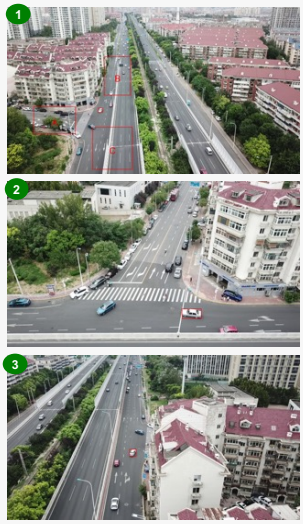
A: Region A
B: Region B
C: Region C



GPT-4o:	C	✗
Qwen3-VL-235B-A22B:	A	✓
Gemini 2.5 Pro:	A	✓
Qwen2.5-VL-72B:	A	✓
Qwen3-VL-32B:	C	✗
InternVL3.5-38B:	B	✗
InternVL3-78B:	C	✗
LLaVA-OneVision-7B:	B	✗
MiniCPM-V-4.5_8B:	C	✗

Q:At present, there are three unmanned aerial vehicles (UAVs), and the shooting angles are shown in the figure. The UAV from the perspective of Figure 2 is the main command angle, while the perspectives of Figure 1 and Figure 3 are the second and third angles, respectively. The first perspective drone has exited tracking due to low battery. In order to maintain the original plan for tracking, which area should a new drone be added in?

A: Region A
B: Region B
C: Region C



GPT-4o:	C	✗
Qwen3-VL-235B-A22B:	C	✗
Gemini 2.5 Pro:	C	✗
Qwen2.5-VL-72B:	C	✗
Qwen3-VL-32B:	C	✗
InternVL3.5-38B:	B	✓
InternVL3-78B:	C	✗
LLaVA-OneVision-7B:	C	✗
MiniCPM-V-4.5_8B:	C	✗

Figure 15. Additional Examples of Sub-Tasks (Part 8).

Ground-Target Planning

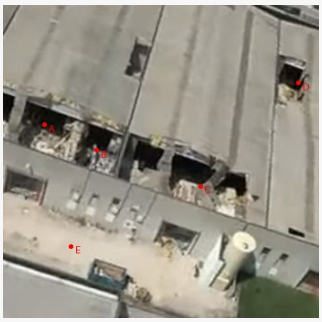
Q:Due to the heavy flood, the rescue team needs to reach the five locations A, B, C, D and E for rescue. Based on water depth, the estimated number of injured or trapped people, and other relevant factors, the rescue priority levels at the locations are as follows: A:High , B:low, C: high, D: low, E: low. Please plan the most reasonable route that prioritizes rescue priority while also minimizing the overall rescue time.

A: E→C→B→A→D
B: E→B→A→C→D
C: E→C→A→D→B
D: E→A→C→B→D

Gemini 2.5 Flash:	/	✗
GPT-4o:	C	✗
Qwen3-VL-235B-A22B:	C	✗
Gemini 2.5 Pro:	C	✗
Qwen2.5-VL-72B:	D	✓
Qwen3-VL-32B:	C	✗
InternVL3.5-38B:	/	✗
InternVL3-78B:	D	✓
LLaVA-OneVision-7B:	C	✗
MiniCPM-V-4.5_8B:	C	✗

Q:After a factory collapse, rescuers should start from point E, rescue the injured at points ABCD, and then take the injured back to point E for evacuation. Considering the rescue principle of prioritizing closer and easier-to-reach victims to avoid secondary collapses, please plan the most suitable rescue route.

A: E→A→B→C→D
Logic: Straight from E to A, along roof residual structure to B, through core to C, finish D.
B: E→D→C→B→A
Logic: Reverse from E to D, clear C and then back to B and A.
C: E→C→A→B→D
Logic: E→C to keep "next to the intact roof", C→A through the core, A→B to clear the residue, B→D along the edge.
D: E→B→A→D→C
Logic: E→B borrows the "residual frame", B→A follows the gentle slope of the collapse, A→D clears the edges, D→C completes, and C returns directly to point E and leaves



Gemini 2.5 Flash:	/	✗
GPT-4o:	D	✗
Qwen3-VL-235B-A22B:	A	✓
Gemini 2.5 Pro:	C	✗
Qwen2.5-VL-72B:	C	✗
Qwen3-VL-32B:	A	✓
InternVL3.5-38B:	C	✗
InternVL3-78B:	A	✓
LLaVA-OneVision-7B:	C	✗
MiniCPM-V-4.5_8B:	D	✗

Figure 16. Additional Examples of Sub-Tasks (Part 9).