Advancing MLLM-based UAV Image Understanding and Reasoning: A Benchmark and a Training-Free Multi-Agent System

Haoyu Zhang, Shuoxun Zhang, Peng Ye, Lin Zhang, Jiakang Yuan, Shenghong Yi, Yuening Wang, Tao Chen, *Senior Member, IEEE*

***Abstract*—Multimodal Large Language Model (MLLM)-based UAV aerial image understanding and reasoning is essential for aerial intelligence yet poses distinct challenges arising from extreme scale variation, arbitrary camera orientations, and high object density. Despite growing interest, existing evaluations remain fragmented across individual datasets and narrow tasks, leaving a critical gap in unified assessment of UAV understanding and reasoning capabilities. To fill this gap, we construct UAVQA-Bench, a benchmark of 1,500 human-annotated QA pairs drawn from 13 public UAV datasets, covering 6 capability dimensions and 16 tasks in both multiple-choice and visual grounding formats. Systematic evaluation of a broad range of open-source and closed-source MLLMs as well as agent-based systems on UAVQA-Bench identifies three key failure modes: domain-toolset mismatch, unchecked error propagation, and static reasoning. Motivated by these findings, we propose UAV-MAS, a training-free multi-agent system for MLLM-based UAV aerial image understanding and reasoning, comprising a Domain-Specific Perception Engine (DSPE) that routes queries to task-appropriate visual tools, a Context-Aware Iterative Refinement module (CAIR) that validates intermediate reasoning to curb error accumulation, and a Difficulty-Aware Adaptive Search mechanism (DAAS) that adjusts search depth to question difficulty. UAV-MAS with a 32B open-source MLLM achieves 77.0% overall accuracy on UAVQA-Bench, surpassing Gemini 3 Pro by 4.0%, while the 8B variant improves 8.7% over its base model.**

***Index Terms*—UAV aerial image analysis, multi-agent systems, multimodal large language models, visual reasoning**

I. Introduction

Unmanned Aerial Vehicles (UAVs) have evolved into indispensable assets across a spectrum of critical applications, from disaster rescue and precision agriculture to urban traffic patrol and infrastructure inspection. In these high-stakes scenarios, the ability to accurately perceive and interpret complex environments is essential for supporting sophisticated decision-making. Although UAVs provide a unique bird's-eye view that enables comprehensive situational awareness, this perspective introduces a formidable set of perceptual challenges: extreme scale variations resulting from altitude fluctuations, arbitrary object orientations inconsistent with ground-level priors, and a prevalence of tiny, densely packed targets amidst cluttered backgrounds. Consequently, advancing UAV visual perception to handle these intricate dynamics is not merely an optimization task but a fundamental prerequisite for reliable aerial artificial intelligence.

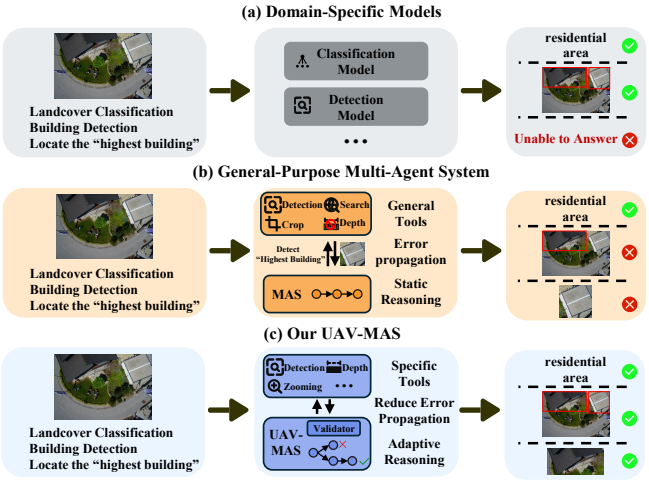


Fig. 1. Comparison of MLLM-based UAV aerial image understanding and reasoning paradigms. (a) **Domain-Specific Models** handle narrow tasks but fail on complex combinatorial queries. (b) **General-Purpose MAS** suffer from toolset domain gaps and error propagation in linear static reasoning. (c) **UAV-MAS (Ours)** addresses all three failure modes via DSPE, CAIR, and DAAS.

Despite significant progress, current UAV perception methods (for example, UAV-DETR [1] for detection and SMDE [2] for depth estimation) remain confined to their respective task domains. As illustrated in Figure 1(a), these task-specific models exhibit two critical limitations when facing comprehensive scenarios. First, they fail to autonomously handle complex combinatorial tasks, as they lack the mechanisms to integrate fragmented visual cues across multi-stage processes such as *"Locating the highest building"*, which requires sequentially binding depth estimation with object detection. Second, they are completely devoid of the cognitive capacity for high-level reasoning and decision-making, rendering them powerless against open-ended problems like *"Assessing if a specific area is suitable for emergency landing"*. Generally speaking,

This work was supported by the National Key R&D Program of China (No. 2026YFE0101200) and the Shanghai Natural Science Foundation (No. 23ZR1402900). The computations in this research were performed using the CFFF platform of Fudan University. Haoyu Zhang, Shuoxun Zhang, and Peng Ye contributed equally to this work. Corresponding author: Tao Chen.

Haoyu Zhang, Shuoxun Zhang, Lin Zhang, Jiakang Yuan and Shenghong Yi are with the Embedded Deep Learning and Visual Analysis Laboratory, College of Future Information Technology, Fudan University, Shanghai 200433, China (e-mail: 25113090081@m.fudan.edu.cn; 24210720324@m.fudan.edu.cn; 22110720068@m.fudan.edu.cn; jkyuan22@m.fudan.edu.cn; 22307130187@m.fudan.edu.cn).

Tao Chen is with the Embedded Deep Learning and Visual Analysis Laboratory, College of Future Information Technology, Fudan University, Shanghai 200433, China, and also with Shanghai Innovation Institute, Shanghai 200231, China (e-mail: eetchen@fudan.edu.cn).

Peng Ye is with the Embedded Deep Learning and Visual Analysis Laboratory, College of Future Information Technology, Fudan University, Shanghai 200433, China, also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China, and also with The Chinese University of Hong Kong, Hong Kong (e-mail: 20110720039@fudan.edu.cn).

Yuening Wang is with the College of Future Information Technology, Fudan University, Shanghai 200433, China (e-mail: 23307130457@m.fudan.edu.cn).

TABLE I
COMPARISON OF UAVQA-BENCH WITH EXISTING BENCHMARKS. **#TASKS** REPRESENTS THE TOTAL NUMBER OF SUBTASKS. **REGION-LEVEL** REFERS TO TASKS WITH PRECISE LOCATION ANNOTATIONS, E.G., BOUNDING BOXES. AND **RELATION** REFERS TO RELATIONS BETWEEN REGION-LEVEL OBJECTS. * DENOTES HUMAN VERIFICATION AFTER MLLM ANNOTATION GENERATION. RULE-BASED DENOTES AUTOMATIC GENERATION FROM STRUCTURED METADATA WITHOUT AI MODELS. UAVQA-BENCH COVERS WIDE TASK SCOPES AND PROVIDES COMPREHENSIVE FULLY HUMAN-ANNOTATED QUESTIONS, OPTIONS, ANSWERS AND GROUNDING BOUNDING BOXES FOR THE EVALUATION OF MLLMS IN THE UAV DOMAIN.

Benchmarks	Domain	Response Format	#Tasks	Scope			Annotation
				Scene-Level	Region-Level	Relation	
MMBench [3]	General	QA	20	✓	✗	✓	Human
RSVQA [4]	Remote Sensing	QA	5	✓	✗	✗	Rule-based
VRSBench [5]	Remote Sensing	QA	12	✓	✓	✗	GPT-4V*
XLRS-Bench [6]	Remote Sensing	QA	16	✓	✓	✓	GPT-4o*
VisDrone [7]	UAV	Bounding Box	4	✗	✓	✗	Human
UAVDT [8]	UAV	Bounding Box	3	✗	✓	✗	Human
UrbanVideo-Bench [9]	UAV	QA	16	✓	✗	✗	Gemini-1.5*
UAVQA-Bench	**UAV**	**QA & Bounding Box**	**16**	✓	✓	✓	**Human**

existing approaches perform well as specialized sensors but lack the holistic 'eyes' and logical 'brain' for complex tasks.

To overcome these limitations, recent Multimodal Large Language Models (MLLMs) offer a promising direction. To advance MLLM-based UAV aerial image understanding and reasoning, we first introduce UAVQA-Bench, a comprehensive and fully human-annotated benchmark. As summarized in Table I, previous comprehensive multimodal benchmarks are largely built on general and remote sensing images, while existing UAV benchmarks still provide only partial coverage of the problem space. In addition, they often rely heavily on automated annotation pipelines, which may compromise data quality. As a comparison, our UAVQA-Bench covers 6 capability dimensions and 16 tasks over 1,500 carefully curated samples drawn from 13 diverse public UAV datasets, and adopts a closed-ended, objectively scorable evaluation protocol that enables reliable and reproducible assessment. All questions, candidate options, answers, and grounding annotations are produced and verified by human annotators, ensuring higher correctness, consistency, and contextual faithfulness.

With the dataset established, a systematic evaluation of a broad range of MLLMs and MLLM-based multi-agent systems is conducted. Initial findings suggest that these methods are generally capable of processing diverse problem types within UAV scenarios, demonstrating an acceptable baseline in general aerial image understanding and reasoning. In detail, closed-source methods show an overall performance advantage. Among open-source methods, larger models generally perform better in aggregate, whereas smaller models can still be stronger on specific tasks such as visual grounding. Despite this functional baseline, our further analysis reveals three recurring failure modes as shown in Figure 1(b): (*i*) **domain-toolset mismatch**, where standard vision tools trained on ground-level data degrade on aerial patterns; (*ii*) **unchecked error propagation**, where early-stage tool errors cascade through multi-step chains without self-correction; and (*iii*) **static reasoning**, where fixed linear strategies fail to adapt to varying aerial task complexity.

To address the above challenges, we propose UAV-MAS, a training-free multi-agent system for MLLM-based UAV aerial image understanding and reasoning. UAV-MAS integrates three targeted components. First, the Domain-Specific Perception Engine (DSPE) equips the system with specialized aerial operators and a precise tool selection mechanism to overcome domain-toolset mismatch on UAV imagery. Second, the Context-Aware Iterative Refinement (CAIR) strategy introduces hierarchical verification that actively detects and corrects reasoning errors, preventing noise accumulation in multi-step inference. Third, the Difficulty-Aware Adaptive Search (DAAS) mechanism dynamically adjusts reasoning level based on task complexity, balancing computational efficiency with thoroughness. Experiments demonstrate strong robustness and reasoning accuracy: UAV-MAS with 32B MLLMs achieves 4.0% higher Overall Accuracy on UAVQA-Bench than the closed-source Gemini 3 Pro, while relying solely on training-free open-source models.

In summary, our main contributions are:

- We introduce UAVQA-Bench, a comprehensive, fully human-annotated benchmark for UAV aerial image understanding and reasoning. It covers 6 capability dimensions across 16 tasks (in multiple-choice and visual grounding formats), with 1,500 samples drawn from 13 diverse public UAV datasets. We systematically evaluate a broad range of open/closed-source MLLM-based methods on UAVQA-Bench.
- We propose UAV-MAS, a training-free multi-agent system for MLLM-based UAV aerial image understanding and reasoning. It integrates a Domain-Specific Perception Engine (DSPE) for aerial-adapted tool use, a Context-Aware Iterative Refinement (CAIR) strategy for error-resilient multi-step reasoning, and a Difficulty-Aware Adaptive Search (DAAS) mechanism for complexity-adaptive inference.
- Extensive experiments show that UAV-MAS achieves state-of-the-art performance, surpassing the powerful closed-source Gemini 3 Pro by 4.0% in Overall Accuracy on UAVQA-Bench using only training-free 32B MLLMs, while offering a more efficient accuracy–cost trade-off than brute-force test-time scaling at comparable accuracy.

II. RELATED WORK

A. UAV Aerial Image Analysis

Recent advances in deep learning have spurred the development of specialized models tailored for diverse analysis tasks in the UAV domain, achieving impressive results. These tasks

include detection [1], [10]–[14], counting [15]–[17], depth estimation [2], [18]–[20]. Representative detection methods include UAV-DETR [1], an end-to-end detector tailored to small objects in drone images; global–local feature fusion with semantic guidance [13]; and detection-oriented exposure correction for nighttime drone views [14]. Recent methods such as FLDet [21] and TGCADNet [22] respectively emphasize lightweight detection and text-guided contextual modeling for small objects in UAV scenes. Although effective under their respective settings, these methods remain optimized for predefined categories and specific perception tasks. Rex-Omni [23] leverages the strong generalization of MLLMs to enable open-world object detection. It reformulates detection as next-point prediction with quantized coordinates and achieves competitive zero-shot performance. For depth estimation, Shi et al. [20] recover scene depth from consecutive UAV frames by combining affine patch transformations with geometric constraints, targeting the narrow-baseline conditions common in aerial sequences.

Despite these advances, existing UAV aerial image analysis models remain confined to isolated tasks and cannot integrate diverse visual capabilities for multi-step reasoning or adaptive decision-making in real-world scenarios.

B. UAV Aerial Image Understanding Benchmarks

Existing benchmarks for evaluating UAV aerial image understanding and reasoning exhibit notable limitations. Typical datasets such as VisDrone [7], [24] and DOTA [25] primarily serve narrow perception primitives like detection and tracking, falling short of assessing higher-order cognitive capabilities. UAVScenes [26] extends to multi-modal perception and six task types including segmentation, yet its focus remains on perceptual primitives rather than comprehensive reasoning. RSVQA [4] provides rule-generated QA for satellite and aerial orthophotos, but does not target low-altitude UAV scenes. More recent remote-sensing benchmarks, such as VRS-Bench [5] and XLRS-Bench [6], remain dominated by nadir-view perspectives and task designs that fail to capture extreme scale variation, oblique viewpoints, and dense spatial layouts.

Recent benchmarks including UrbanVideo-Bench [9] have attempted to construct more comprehensive evaluation frameworks. However, they generally suffer from incomplete task coverage and incompatible settings, along with reliance on automated annotation pipelines, resulting in data quality that falls short of human-annotated standards.

C. Multi-Agent System

Multi-agent systems powered by MLLMs have recently gained traction as a framework for tackling complex, multi-step visual reasoning tasks, where collaborative agents decompose high-level queries and coordinate visual analysis with reasoning [27]–[36]. ReAct [37] introduces a prompting framework that interleaves reasoning and actions, enabling language models to reason and act jointly. It can reduce hallucination and improve performance in tasks like question answering and decision-making. DyFo [27] adopts a static toolkit architecture and builds a multimodal interaction framework based on ReAct, and utilizes the MCTS search algorithm to achieve a method that enhances the fine-grained visual understanding capability of large-scale multimodal models without training. PyVision [28] adopts the Model Context Protocol (MCP) toolkit as the interaction interface, follows the ReAct paradigm to implement iterative cycles of multi-step reasoning and tool invocation, and utilizes the MCTS search algorithm to enhance the adaptability and interpretability of complex visual reasoning tasks.

Although effective in their domains, these methods are ill-suited for UAV aerial image understanding due to ground-biased vision tools, error propagation from tool chaining, and rigid planning that fails to adapt to aerial task complexity—motivating the need for a resilient and dynamically adaptive multi-agent framework tailored to the aerial domain.

III. UAVQA-BENCH

To address the limitations in existing UAV aerial image understanding and reasoning evaluation platforms, we introduce "UAVQA-Bench", a comprehensive benchmark covering multi-task, multi-scenario, and multi-scale settings. As shown in table I, UAVQA-Bench features comprehensive scopes including scene-level, region-level and relation between regional objects, which ensures a holistic evaluation beyond the scope of existing UAV datasets. All annotations are provided by highly educated human annotators under an inspection-revision process. Compared to synthetic annotations in benchmarks like VRS-Bench [5] and UrbanVideo-Bench [9], human-crafted annotations demonstrate superior quality in terms of correctness, consistency, and contextual understanding over synthetic outputs from large models. UAVQA-Bench assesses 6 key capabilities of UAV agent systems and encompasses 16 distinct tasks, systematically examining UAV agents' multi-dimensional understanding and reasoning abilities in complex environments. Notably, our evaluation adopts a closed-ended QA (multiple-choice) and grounding box matching approach, which simplifies assessment and enhances reproducibility by eliminating the ambiguity inherent in open-ended responses, while ensuring reliable evaluation through standardized answer spaces.

A. Dataset Construction and Data Sources

UAVQA-Bench comprises 1,500 samples, each pairing a high-quality UAV image with a carefully constructed question–answer instance. To ensure task relevance and visual diversity, we manually selected images from 13 publicly available UAV datasets: AU-AIR [38], WebUAV-3M [39], VisDrone-DET2019 [40], Semantic Drone [41], DroneVehicle [42], UAVDT [8], VDD [43], UDD [44], UAVid [45], WildUAV [46], HazyDet [47], AnimalDrone [48], and UAV123 [49], with their task-level provenance summarized in table II. Dataset construction followed a multi-stage quality-control process involving seven volunteers with backgrounds in computer vision: two selected suitable images and annotated bounding boxes where necessary, two constructed the questions, answer options, and ground-truth answers, and

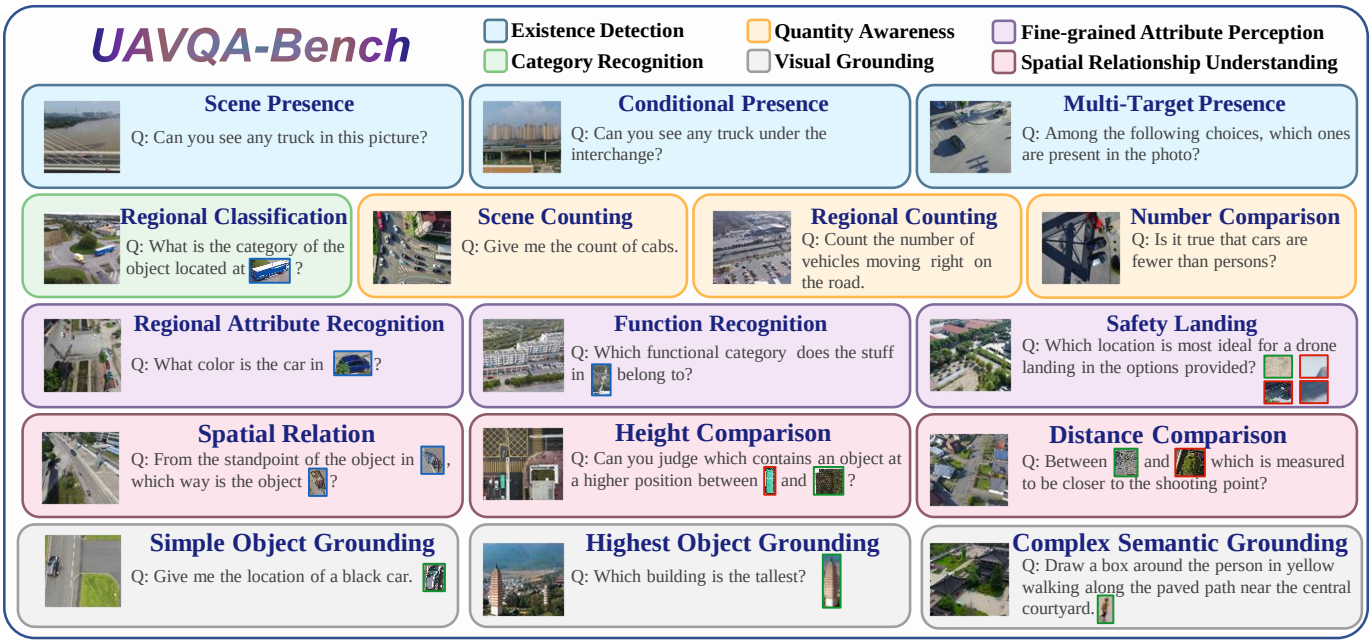


Fig. 2. Overview of UAVQA-Bench. The benchmark assesses 6 key capabilities through 16 distinct tasks, collectively forming a comprehensive evaluation framework.

TABLE II
DATA SOURCES OF EACH TASK.

Capability	Task	Data Source
Existence Detection	Scene Presence	WebUAV-3M, VisDrone-DET2019, Semantic Drone
	Conditional Presence	AU-AIR, WebUAV-3M, Semantic Drone
	Multi-Target Presence	AU-AIR, WebUAV-3M, VisDrone-DET2019, Semantic Drone, UAVDT, HazyDet
Category Recognition	Regional Classification	AU-AIR, VisDrone-DET2019
Quantity Awareness	Scene Counting	AU-AIR, WebUAV-3M, Semantic Drone, AnimalDrone
	Regional Counting	AU-AIR, WebUAV-3M, VisDrone-DET2019, Semantic Drone, AnimalDrone
	Number Comparison	AU-AIR, WebUAV-3M, VisDrone-DET2019, Semantic Drone, UAVDT, HazyDet, AnimalDrone, UAV123
Fine-grained Attribute Perception	Regional Attribute Recognition	AU-AIR, VisDrone-DET2019, DroneVehicle
	Function Recognition	VisDrone-DET2019, VDD, UDD
	Safety Landing	VisDrone-DET2019
Spatial Relationship Understanding	Spatial Relation	AU-AIR, VisDrone-DET2019, Semantic Drone
	Height Comparison	VisDrone-DET2019, UAVDT, VDD, UDD, UAV123
	Distance Comparison	AU-AIR, VisDrone-DET2019, UAVDT, VDD, UDD, UAVid, WildUAV
Visual Grounding	Simple Object Grounding	AU-AIR, VisDrone-DET2019
	Complex Semantic Grounding	AU-AIR, WebUAV-3M
	Highest Object Grounding	WebUAV-3M, VDD, Semantic Drone

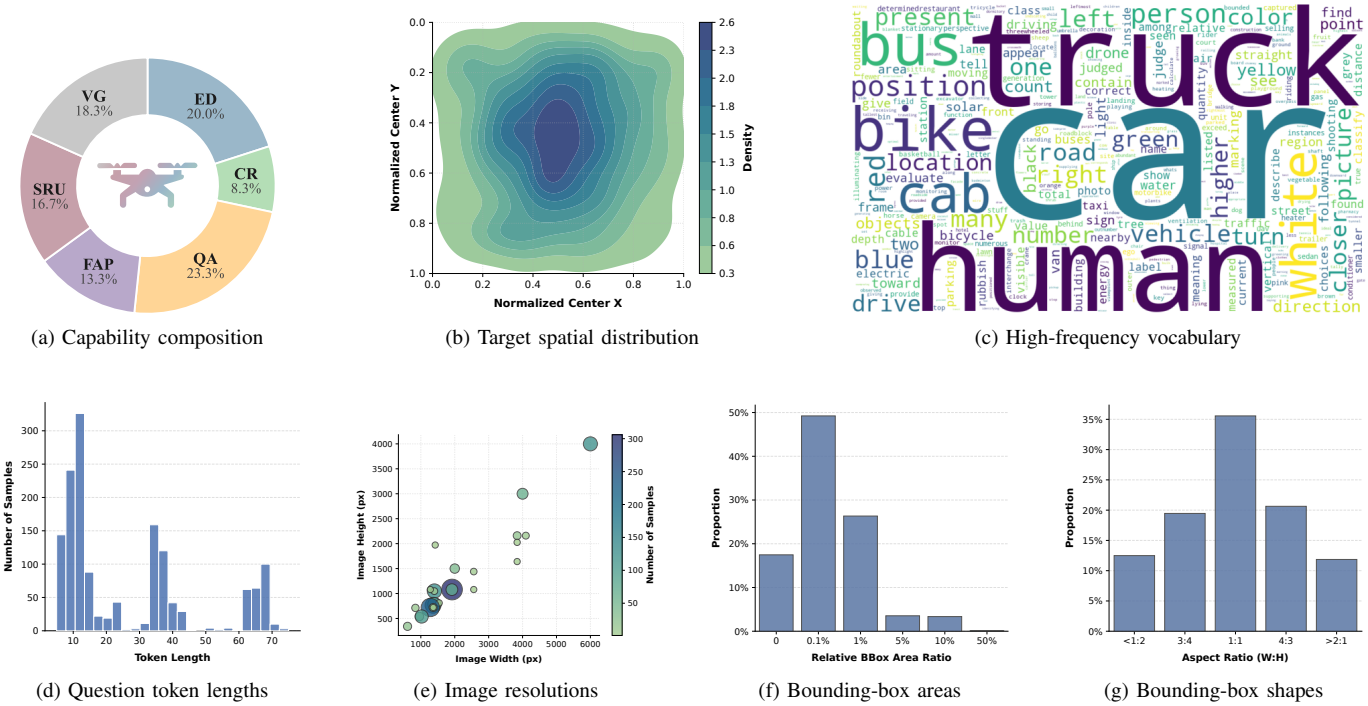
the remaining three independently reviewed the image quality, bounding-box annotations, question and option quality, and answer correctness. For tasks requiring distractor options, three distractors were selected from six LLM-generated candidates. Only samples receiving unanimous approval were retained; the remaining samples were revised or removed, followed by random spot checks of the resulting dataset.

For region-level tasks, we manually annotate the selected images with precise bounding boxes. We then design task-specific question templates with diverse linguistic expressions to reduce the risk of models exploiting superficial patterns. Candidate answers for multiple-choice questions include plausible visual or semantic distractors, while all questions, options, ground-truth answers, and grounding annotations are manually authored and rigorously cross-verified. For tasks requiring spatial outputs, such as visual grounding, coordinates are represented as $\{<x_1><y_1><x_2><y_2>\}$, where (x_1, y_1) and (x_2, y_2) denote the top-left and bottom-right corners, respectively. All coordinates are normalized by image width and height and rounded to three decimal places for consistent spatial representation.

B. Capability Dimensions

As illustrated in Figure 2, UAVQA-Bench establishes 6 core capabilities comprising 16 diverse tasks, ranging from basic perception to high-level reasoning.

- **Existence Detection (ED)** evaluates the agent's ability to identify object presence under varying conditions, including *Scene Presence*, *Conditional Presence*, and *Multi-Target Presence*.
- **Category Recognition (CR)** focuses on fundamental semantics through *Regional Classification* of detected entities.



(a) Capability composition (b) Target spatial distribution (c) High-frequency vocabulary

(d) Question token lengths (e) Image resolutions (f) Bounding-box areas (g) Bounding-box shapes

Fig. 3. Dataset statistics of UAVQA-Bench. (a) Composition of the six capability dimensions. (b) Spatial distribution of target centers. (c) High-frequency vocabulary. (d) Question token-length distribution. (e) Image-resolution distribution. (f) Bounding-box area relative to image area. (g) Bounding-box aspect-ratio distribution.

- **Quantity Awareness (QA)** needs precise counting and numerical logic via *Scene/Regional Counting* and *Number Comparison*.
- **Fine-grained Attribute Perception (FAP)** assesses the agent's granular understanding of entity properties and scene affordances. It encompasses *Attribute Recognition* (e.g., identifying colors and orientations), *Function Recognition* of specific objects, and *Safety Landing* analysis, which requires the model to evaluate surface conditions to identify secure touchdown zones for the UAV.
- **Spatial Relationship Understanding (SRU)** examines 3D spatial reasoning from a distinctive aerial perspective. Beyond identifying *Spatial Relations* between discrete entities or their relative positioning to the UAV, this dimension includes *Height Comparison* and *Distance Comparison*. Here, "higher" denotes greater physical elevation in 3D space rather than a smaller image-plane vertical coordinate, while distance denotes the camera-to-object range. Agents must therefore reason about relative elevation and depth in complex, multi-level environments.
- **Visual Grounding (VG)** tests the precise alignment between complex linguistic queries and visual regions. Spanning from *Simple Object Grounding* to *Complex Semantic Grounding* with multi-step reasoning, this capability evaluates the agent's cross-modal comprehension. Notably, it includes *Highest Object Grounding*, which challenges the model to identify and localize the vertically most prominent entities, such as the tallest building in a dense urban scene.

C. Evaluation Protocol and Metrics

Most UAVQA-Bench tasks adopt a multiple-choice format, whereas visual grounding tasks require a predicted bounding box. A multiple-choice instance is counted as correct when the model selects the ground-truth option. For visual grounding, the model is explicitly prompted to return coordinates in the same normalized format used by the annotations, and a prediction is counted as correct when its Intersection over Union (IoU) with the ground-truth box reaches or exceeds 0.5.

We report two complementary aggregate metrics: **Overall Accuracy (OA)** and **Average Accuracy (AA)**. Let $T = 16$ denote the number of tasks, and let m_i and c_i denote the total and correctly answered samples in task i, respectively. The metrics are defined as

$$\mathrm{OA} = \frac{\sum_{i=1}^{T} c_i}{\sum_{i=1}^{T} m_i} \times 100\%, \tag{1}$$

$$\mathrm{AA} = \frac{1}{T} \sum_{i=1}^{T} \frac{c_i}{m_i} \times 100\%. \tag{2}$$

OA measures sample-weighted performance over the complete benchmark, while AA assigns equal weight to every task. Reporting both metrics therefore captures overall robustness without allowing tasks with more samples to obscure performance differences across capability dimensions.

D. Dataset Statistics

Figure 3 provides a unified statistical overview of UAVQA-Bench. The benchmark maintains a relatively balanced com-

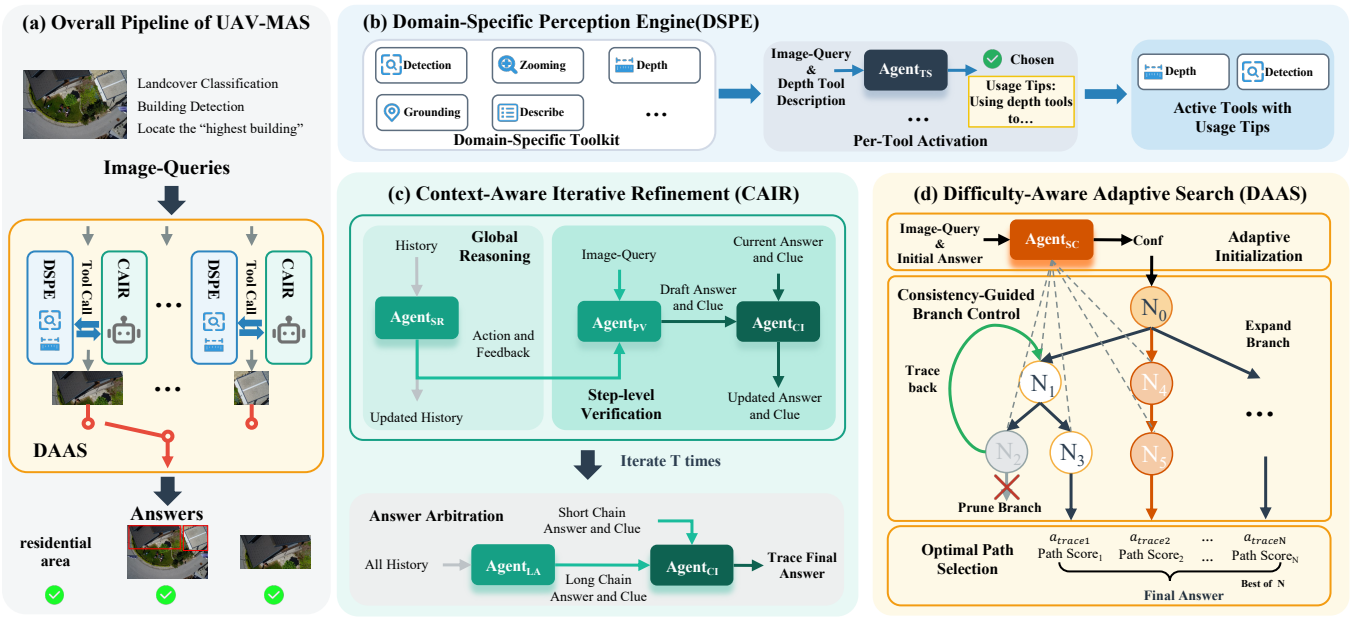


Fig. 4. Overview of UAV-MAS. (a) The overall pipeline of UAV-MAS. (b) The Domain-Specific Perception Engine (DSPE) tailored for UAV perception. (c) The Context-Aware Iterative Refinement (CAIR) strategy to mitigate tool invocation noise. (d) The Difficulty-Aware Adaptive Search (DAAS) mechanism for dynamic reasoning exploration.

position across the six capability dimensions (Figure 3(a)), covering both foundational perception and more challenging reasoning-oriented tasks. Target objects are broadly distributed across the image plane rather than concentrated in a small set of locations (Figure 3(b)). The word cloud in Figure 3(c) further summarizes the benchmark's lexical coverage across object categories, attributes, spatial relationships, and actions. Non-content words, such as prepositions and formatting tokens, are removed so that font size directly reflects the relative frequency of semantically meaningful terms.

UAVQA-Bench also presents substantial diversity in both textual and visual inputs: question lengths range from short queries to long compositional reasoning prompts, while image resolutions span from low-resolution to ultra-high-resolution aerial imagery (Figure 3(d) and (e)). The bounding-box statistics in Figure 3(f) and (g) further reveal the fine-grained perception challenge. Specifically, 49.2% of the boxes occupy only 0.1%–1% of the image area, and another 17.5% occupy less than 0.1%; by contrast, just 0.1% cover more than 50% of an image. Near-square boxes, with width-to-height ratios between 3:4 and 4:3, form the largest group, although the complete distribution spans a wide range of object shapes. Together, these statistics demonstrate diversity in capability coverage, target location and geometry, language, and image resolution, making UAVQA-Bench a challenging testbed for multimodal agents in UAV scenarios.

IV. Method

Evaluating MLLM-based methods on UAVQA-Bench (Table III) exposes three key failures: domain-toolset mismatch, error propagation, and static reasoning. To address these, we propose **UAV-MAS** (Figure 4), a training-free multi-agent system comprising three core modules. First, the **Domain-Specific Perception Engine (DSPE)** provides aerial-targeted tools to overcome domain mismatch. Second, **Context-Aware Iterative Refinement (CAIR)** utilizes step-level verification to correct unreliable feedback and halt error propagation. Third, **Difficulty-Aware Adaptive Search (DAAS)** replaces static reasoning with complexity-adaptive exploration by pruning paths based on query difficulty. Together, these modules enable **UAV-MAS** to achieve robust aerial image understanding and reasoning performance without task-specific training.

A. Domain-Specific Perception Engine

While general-purpose Multimodal Large Language Models (MLLMs) generalize well to common scenes, aerial images in UAV understanding and reasoning tasks expose structural challenges that standard models cannot adequately handle: extreme scale variation, arbitrary object orientation, and high-density small object clutter. Generic models also lack priors for aerial-specific reasoning subtasks (e.g., distinguishing ground-level from rooftop targets, counting tiny objects under dense occlusion). To address these gaps, we design the **Domain-Specific Perception Engine (DSPE)**: a modular suite of aerial-targeted tools combined with a per-tool activation mechanism for lightweight scheduling.

Domain-Specific Toolkit. The toolkit comprises 5 core tools targeting the principal visual bottlenecks in overhead imagery: **1) Context-Aware Zooming** employs a soft-boundary mechanism that retains edge context during cropping to enhance local resolution while preserving semantic continuity, particularly important for aerial targets that often span only a few pixels at typical UAV altitudes. **2) Fine-Grained Explicit Description** converts implicit visual features (e.g., textures,

behaviors) into structured text via instruction following, supporting downstream reasoning. **3) Distance Estimation** utilizes Depth Anything 3 [50] to generate pseudo-depth maps for reconstructing 3D structures from monocular imagery, distinguishing ground-level from rooftop targets based on relative depth discontinuities. **4) Semantic Grounding** locates individual targets based on explicit instructions (e.g., "Locate the red vehicle") for precise spatial positioning. **5) Open-Vocabulary Detection with De-hallucination** addresses hallucination patterns specific to dense aerial scenes through a two-stage verification. First, prior to localizing, an MLLM-based existence verification confirms whether the target category is present at all, preventing the model from forcibly generating bounding boxes for absent targets. Second, to address the issue where MLLMs systematically hallucinate bounding boxes at regular intervals during autoregressive decoding, we apply a heuristic filter. It removes a detection c_i only when the equidistant condition $|c_{i+1}-2c_i+c_{i-1}| < \delta$ holds *continuously* across multiple consecutive detections, preserving legitimate near-uniform layouts (e.g., a row of parked vehicles) while suppressing fabricated repetitive patterns.

Per-Tool Activation. In aerial image analysis, tool selection is query-type-dependent: counting tasks require detection while spatial reasoning tasks require depth estimation rather than description. Small open-source MLLMs frequently fail when asked to jointly select and configure multiple tools from a single prompt, due to context overflow or hallucinated activations. To address this, each tool is assigned a dedicated lightweight agent (Agent_{TS}) responsible for a single binary decision: "Should this tool be activated for the current image-query pair?" If activated, Agent_{TS} additionally generates the required execution parameters.

Together, the 5 aerial-targeted tools and per-tool activation ensure that the base MLLM receives enriched, task-relevant perceptual inputs without needing to manage complex multi-tool scheduling—directly addressing the domain gap in UAV aerial image understanding and reasoning.

B. Context-Aware Iterative Refinement (CAIR)

In aerial image QA, tool feedback (e.g., from object detectors) is inherently noisy due to occlusion, degraded image quality, and the small scale of targets in overhead imagery. To prevent such errors from corrupting the accumulated reasoning state, we propose the **Context-Aware Iterative Refinement (CAIR)** strategy. CAIR augments the standard ReAct loop with a step-level verification stage: after each tool call, a dedicated agent assesses whether the new evidence is reliable and consistent, and selectively updates the answer and accumulated key clues accordingly.

Global Reasoning Track. The **Strategic Reasoning Agent (Agent_{SR})** maintains a ReAct [37] loop over the dialogue history $\mathcal{H}_{t-1}$ of the current trace, selecting actions A_t from the tool set provided by DSPE and accumulating tool feedback F_t:

$$\mathcal{H}_t = \mathcal{H}_{t-1} || \{A_t, F_t\}. \tag{3}$$

Step-level Verification. At each reasoning step, the **Perceptual Verification Agent (Agent_{PV})** receives only the current step's data: image I, query Q, action A_t, and tool feedback F_t, without access to the accumulated history. This isolation prevents historical anchoring bias and produces a draft answer $\hat{a}_t$ and a compact evidence summary $\hat{C}_t$. The **Contextual Integration Agent (Agent_{CI})** does not receive the raw tool output F_t; instead, it compares $(\hat{a}_t, \hat{C}_t)$ with the original image–query context and the previous state (a_{t-1}, C_{t-1}). The image I and query Q are persistent shared context and are omitted from repeated agent-call notation for brevity. The resulting state update is

$$(a_t,\, C_t) = \begin{cases} (\hat{a}_t,\, \hat{C}_t) & \text{if trusted and } \hat{a}_t \neq a_{t-1}, \\ (a_{t-1},\, C_{t-1} || \hat{C}_t) & \text{if trusted and } \hat{a}_t = a_{t-1}, \\ (a_{t-1},\, C_{t-1}) & \text{otherwise,} \end{cases} \tag{4}$$

where $||$ denotes text-level concatenation, a_t is the current best answer, and C_t is a structured text record of accumulated key clues. Because the extracted clues are tightly coupled with the specific answer hypothesis, the state is fully replaced when the draft answer changes to avoid evidence contamination between conflicting hypotheses; when it remains consistent, new clues are safely appended to accumulate evidence.

Answer Arbitration. Upon completion of the reasoning loop, the **Long-Chain Answer Agent (Agent_{LA})** synthesizes an answer a_{end} and a clue summary C_{end} from the accumulated history $\mathcal{H}_T$ of that trace. The **Contextual Integration Agent (Agent_{CI})** then selects the more reliable result between a_{end} and the final iterative state a_T as the trace output a_{trace}.

CAIR's step-level verification protects the answer–evidence state within each trace. During DAAS, every child receives a copy of its parent's state and maintains an independent history, so feedback on an unreliable or pruned branch never enters a sibling branch. The complete procedure is summarized in Algorithm 1.

Algorithm 1 Context-Aware Iterative Refinement (CAIR)

Input: Task Query Q, Image I, Iteration Limit T
Output: Single Trace Final Answer a_{trace}
Initialize: $\mathcal{H}_0 \leftarrow \{Q, I\}$, $a_0 \leftarrow \text{VLM}(Q, I)$
  $C_0 \leftarrow \emptyset$, $\mathcal{T}_{\text{opt}} \leftarrow \text{ToolSelection}(Q, I, \mathcal{T})$
for $t = 1$ **to** T **do**
  Phase 1: Global Reasoning (ReAct)
    $A_t \leftarrow \text{Agent}_{\text{SR}}(\mathcal{H}_{t-1}, \mathcal{T}_{\text{opt}})$
    $F_t \leftarrow \text{Execute}(A_t)$
    Get $\mathcal{H}_t$ based on Eq. 3
  Phase 2: Step-level Verification
    $\hat{a}_t, \hat{C}_t \leftarrow \text{Agent}_{\text{PV}}(I, Q, A_t, F_t)$
    $(a_t, C_t) \leftarrow \text{Agent}_{\text{CI}}(\hat{a}_t, \hat{C}_t, a_{t-1}, C_{t-1})$ *via Eq. 4*
end for
Phase 3: Answer Arbitration
$a_{\text{end}}, C_{\text{end}} \leftarrow \text{Agent}_{\text{LA}}(\mathcal{H}_T)$
$a_{\text{trace}} \leftarrow \text{Agent}_{\text{CI}}(a_{\text{end}}, C_{\text{end}}, a_T, C_T)$
Return a_{trace}

C. Difficulty-Aware Adaptive Search (DAAS)

While CAIR mitigates noise at each reasoning step, it follows a fixed linear chain that may miss better reasoning

TABLE III
COMPARISON WITH OTHER METHODS ON OUR UAVQA-BENCH. GEN. AND SPEC. DENOTE THE USE OF THE ORIGINAL GENERAL-PURPOSE TOOLKIT AND OUR DOMAIN-SPECIFIC TOOLKIT, RESPECTIVELY, BOTH EQUIPPED WITH QWEN3-VL AS THE BACKBONE. OA AND AA DENOTE OVERALL ACCURACY AND AVERAGE ACCURACY, RESPECTIVELY. THE GLM4.6V RESULT UNDER UAV-MAS IS A SUPPLEMENTAL BACKBONE TEST. HUMAN AVG. REPORTS THE AVERAGE PERFORMANCE OF 10 PARTICIPANTS. ALL VALUES ARE REPORTED IN PERCENTAGES (%). **BOLD** INDICATES THE BEST MODEL PERFORMANCE, EXCLUDING THE HUMAN BASELINE.

Category	Method	Model	OA	AA	ED	CR	QA	FAP	SRU	VG
Human Avg. (10 participants)			87.87	86.14	84.33	90.40	85.43	92.50	80.80	83.39
Open Source VLM	*Qwen3-VL [51]*	Qwen3-VL 8B Instruct	61.73	59.63	72.33	52.80	64.57	46.50	41.60	80.00
		Qwen3-VL 8B Thinking	52.60	50.88	79.00	50.40	64.86	44.50	39.60	26.91
		Qwen3-VL 30BA3B Instruct	61.13	59.72	75.00	54.40	58.57	51.50	43.20	75.64
		Qwen3-VL 30BA3B Thinking	62.00	60.03	78.33	52.00	69.43	53.50	56.00	50.91
		Qwen3-VL 32B Instruct	69.60	68.63	80.00	67.20	67.71	58.00	59.60	79.27
		Qwen3-VL 32B Thinking	68.27	67.06	77.33	63.20	67.14	57.00	58.40	79.27
	InternVL3.5 [32]	InternVL3.5 8B	40.53	39.18	58.00	36.00	53.71	36.50	47.20	3.63
		InternVL3.5 38B	55.13	55.73	77.33	66.40	61.14	60.50	51.20	17.82
	GLM4.6V [52]	GLM4.6V 9B	65.27	63.57	72.00	57.60	67.71	51.00	56.40	76.73
	Qwen3.5 [53]	Qwen3.5 9B	63.80	62.81	65.67	59.20	65.14	57.5	50.80	78.55
Closed Source VLM	*ChatGPT [54]*	ChatGPT 5.2 Pro	57.13	56.72	77.67	60.80	71.43	58.50	67.20	4.73
		ChatGPT 5.2	53.20	52.30	75.67	57.60	69.43	46.50	58.80	5.82
	Gemini [55]	Gemini 3 Pro	73.00	71.25	79.67	60.80	76.29	68.50	60.80	81.45
		Gemini 3 Flash	70.73	68.78	**82.33**	59.20	71.71	60.50	56.40	82.55
Agent Based Method	*DyFo [27]*	Qwen3-VL 8B	46.47	48.31	61.00	64.00	48.86	63.00	35.20	17.82
		Qwen3-VL 32B	51.67	52.81	70.00	64.00	57.14	61.50	58.40	5.82
	Qwen-Agent [56]	Qwen3-VL 8B-Gen.	60.87	59.11	75.67	52.00	61.43	52.00	41.20	72.36
		Qwen3-VL 8B-Spec.	61.13	60.46	71.00	60.00	56.57	50.50	51.60	73.09
		Qwen3-VL 32B-Gen.	71.13	71.33	74.33	78.40	77.43	72.50	58.40	66.91
		Qwen3-VL 32B-Spec.	69.40	69.63	63.67	71.20	71.43	71.50	68.00	72.00
	PyVision [28]	Qwen3-VL 8B-Gen.	44.80	44.20	64.00	48.80	63.14	51.50	35.20	2.55
		Qwen3-VL 8B-Spec.	50.87	49.16	62.67	45.60	63.71	44.50	39.20	39.27
		Qwen3-VL 32B-Gen.	50.60	51.74	63.67	68.00	69.71	67.50	41.20	0.36
		Qwen3-VL 32B-Spec.	52.33	52.53	62.67	58.40	61.43	59.00	46.40	27.27
Ours	*UAV-MAS*	**Qwen3-VL 8B**	70.47	68.94	77.67	61.60	69.71	57.50	66.80	80.36
		Qwen3-VL 32B	**77.00**	**76.03**	77.00	**70.40**	**80.57**	**75.00**	69.20	**84.00**
		GLM4.6V 9B	69.67	67.67	75.67	56.80	70.86	54.00	**71.60**	77.09
		Qwen3.5 9B	73.67	73.81	74.00	70.40	74.57	70.50	67.20	81.81

paths for complex aerial queries. In UAV visual question answering, query difficulty varies substantially: a simple presence check requires few reasoning steps, while a compositional query (e.g., counting objects satisfying multiple spatial and attribute conditions) demands deeper, multi-step exploration. Applying a fixed pruning threshold across all queries wastes computation on easy queries and prematurely cuts off exploration on hard ones. To address this, we propose the **Difficulty-Aware Adaptive Search (DAAS)** strategy. Unlike standard beam search, which applies a single fixed pruning threshold and selects the final-node answer, DAAS derives a per-query pruning threshold from estimated query difficulty and selects the optimal path by global coherence across all nodes. DAAS governs tree expansion through three phases: Adaptive Initialization, Consistency-Guided Branch Control, and Optimal Path Selection.

Adaptive Initialization. Before searching, the **Score Agent ($\mathbf{Agent_{SC}}$)** assigns the base model's direct response an initial score $S_{\text{init}} \in [0, 10]$ as a coarse difficulty indicator. We use a fixed mapping: $S_{\text{init}} \in [0, 3]$, $[4, 8]$, and $[9, 10]$ correspond to $\tau = 2$, 4, and 6, respectively. Larger thresholds avoid unnecessary expansion for high-confidence queries, while smaller thresholds preserve exploration for low-confidence queries.

Consistency-Guided Branch Control. We manage tree expansion through a Consecutive Consistency check. During expansion, the system generates W candidate successor states via CAIR using temperature-based sampling, where W controls branching width. The **Score Agent ($\mathbf{Agent_{SC}}$)** evaluates each candidate's quality to output a new score S'. Branch continuation is decided by coherence over consecutive steps:

- **Prune Branch:** We prune only if both the current node and the new candidate fall below the threshold simultaneously:

$$\text{Prune} \iff (S' < \tau) \wedge (S < \tau). \tag{5}$$

This allows a single low-confidence step to persist if its adjacent step remains high-confidence. The key insight is that intermediate tool calls in aerial reasoning (e.g., a zoom step returning a blurry sub-patch before a subsequent description step integrates the result) may

transiently score low without indicating a dead branch. Requiring two *consecutive* low-confidence steps before pruning reduces false terminations while discarding genuinely unproductive paths.

- **Expand Branch:** If at least one step in the consecutive pair maintains high confidence, we create a new node N' and continue exploration from it.

Optimal Path Selection. Tree expansion follows branch-local recursion and terminates when a valid final answer is generated or the maximum depth is reached. A final answer is valid only if it matches the required task format, namely a candidate option for multiple-choice tasks or a parseable normalized box for visual grounding. Paths that reach the depth limit without a valid answer are discarded. Upon completion, we select the optimal valid path by evaluating global coherence across all nodes rather than relying solely on the final-node score. To mitigate path-length bias commonly associated with average scoring across long chains, Agent_{SC} is explicitly prompted to assign strict penalties to uninformative or meandering intermediate steps. The optimal path $\mathcal{P}^*$ is thus selected by maximizing the average node score:

$$\mathcal{P}^* = \arg\max_{\mathcal{P}_i} \left(\frac{1}{|\mathcal{P}_i|} \sum_{N \in \mathcal{P}_i} S(N) \right). \tag{6}$$

Subsequently, the valid answer stored at the leaf of $\mathcal{P}^*$ is selected as the global final answer a_{final}:

$$a_{\text{final}} \leftarrow a_{\text{leaf}}(\mathcal{P}^*). \tag{7}$$

If every branch is pruned or reaches the depth limit without a valid answer, DAAS returns the base model's initial response a_0 as a conservative fallback. In summary, DAAS focuses computation on aerial queries that require multi-step spatial or attribute reasoning by deriving a per-query pruning threshold and selecting the valid path with the highest global coherence. The complete procedure is summarized in Algorithm 2. Implementation details and agent configurations are provided in the supplementary material.

Algorithm 2 Difficulty-Aware Adaptive Search (DAAS)

Input: Image I, Query Q, Max Depth D, Max Width W
Output: Final Answer a_{final}
$a_0 \leftarrow \text{VLM}(Q, I)$, $S_{\text{init}} \leftarrow \text{Agent}_{\text{SC}}(Q, I, a_0)$
$\tau \leftarrow g(S_{\text{init}})$ using the fixed three-level map
$N_{\text{root}} \leftarrow \{\mathcal{H}_0, a_0, C_0, S_{\text{init}}, d = 0\}$, $\mathcal{V} \leftarrow \emptyset$
Function Search($N\{\mathcal{H}, a, C, S, d\}$)
for $k = 1$ **to** W **do**
  Sample one CAIR step to obtain $\mathcal{H}', a', C'$
  $S' \leftarrow \text{Agent}_{\text{SC}}(\mathcal{H}' - \mathcal{H})$
  if $S' < \tau \wedge S < \tau$ **then**
    continue *// prune two consecutive weak steps*
  end if
  $N' \leftarrow \{\mathcal{H}', a', C', S', d + 1\}$
  if $\text{Final}(a') \wedge \text{Valid}(a')$ **then**
    $\mathcal{V} \leftarrow \mathcal{V} \cup \{\mathcal{P}(N')\}$
  else if $d + 1 = D$ **then**
    if $\text{Valid}(a')$ **then**
      $\mathcal{V} \leftarrow \mathcal{V} \cup \{\mathcal{P}(N')\}$
    end if
  else
    Search(N')
  end if
end for
End Function
Search(N_{root})
if $\mathcal{V} = \emptyset$ **then**
  $a_{\text{final}} \leftarrow a_0$ *// fallback for pruned/invalid searches*
else
  Select $\mathcal{P}^*$ from $\mathcal{V}$ using Eq. 6
  $a_{\text{final}} \leftarrow a_{\text{leaf}}(\mathcal{P}^*)$ via Eq. 7
end if
Return a_{final}

V. Experiment

A. Main Results

The quantitative results on UAVQA-Bench are presented in Table III. Built upon Qwen3-VL 8B and 32B models, our proposed UAV-MAS demonstrates substantial performance gains over both vanilla baselines and existing agent frameworks. Specifically, UAV-MAS-8B and UAV-MAS-32B surpass their respective Instruct counterparts by margins of 8.7% and 7.4% in Overall Accuracy. This substantial improvement underscores the efficacy of our design in bridging the domain gap that limits standard MLLMs. Furthermore, compared to general-purpose multi-agent frameworks utilizing identical models, UAV-MAS achieves a 5.9% lead over the strongest baseline, validating the effectiveness of our designs. Remarkably, UAV-MAS-32B outperforms the closed-source Gemini 3 Pro by 4.0%, showing that a domain-specialized open-source multi-agent design can compete effectively with proprietary general-purpose models.

TABLE IV
Module-level ablation of UAV-MAS on UAVQA-Bench. Baseline: Qwen3-VL 8B Instruct. OA: Overall Accuracy (%).

Method	DSPE	CAIR	DAAS	OA
Baseline				61.73
Exp1	✓			64.53
Exp2	✓	✓		68.03
Exp3	✓		✓	68.27
Full	✓	✓	✓	**70.47**

B. Ablation Study

1) Module-Level Ablation: Table IV presents an incremental analysis of each module. Starting from the vanilla Qwen3-VL 8B baseline (61.73%), equipping it with DSPE alone (Exp1) raises accuracy to 64.53%, confirming the necessity of domain-specific perception operators. When CAIR is further added (Exp2), performance jumps substantially to 68.03%, reflecting its effectiveness in suppressing error propagation across reasoning steps. Adding DAAS instead of CAIR (Exp3) also yields a strong gain to 68.27%, demonstrating that adaptive search independently contributes to accurate reasoning.

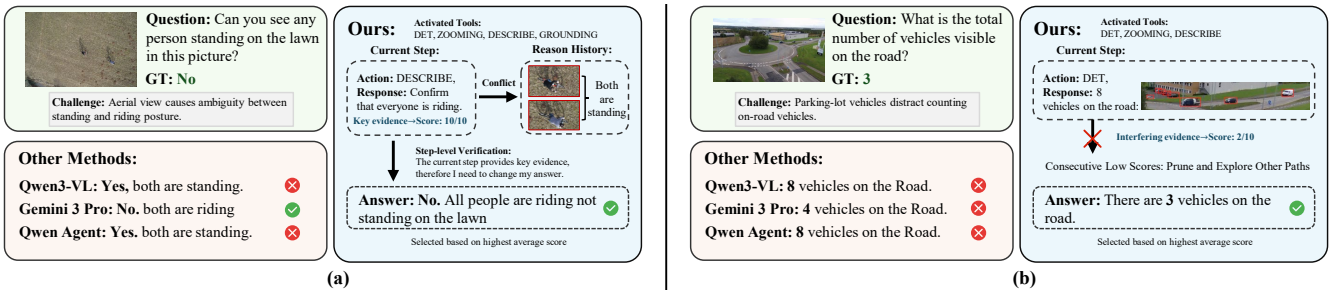


Fig. 5. Visual comparison of UAV-MAS with other methods. All experiments are based on Qwen3-VL 8B except Gemini 3 Pro.

TABLE V
WITHIN-MODULE DESIGN ABLATION OF UAV-MAS. BASELINE: QWEN3-VL 8B INSTRUCT. OA: OVERALL ACCURACY (%).

Module	Variant	OA (%)
Baseline	-	61.73
DSPE	w/o Per-Tool Activation	69.60
	w/o Distributed Agent_{TS}	68.07
CAIR	w/o CAIR	68.27
	w/o Agent_{CI}	70.27
	w/o Agent_{PV}	69.53
DAAS	w/o DAAS	68.03
	w/o Difficulty Awareness	69.87
	w/o Pruning	68.73
Full	-	**70.47**

Integrating all three modules achieves the highest score of 70.47%, indicating that DSPE, CAIR, and DAAS address complementary bottlenecks and together produce a synergistic effect.

2) Within-Module Design Ablation: Table V examines the internal design choices of each module against the full system (70.47%). For DSPE, disabling Per-Tool Activation—which feeds all available tools directly into the agent—causes accuracy to drop to 69.60%, highlighting the critical importance of actively selecting task-relevant tools. Furthermore, collapsing the parallel, distributed Agent_{TS} into a single centralized agent degrades performance to 68.07%, demonstrating that in domain-specific scenarios, smaller, less capable models require highly fine-grained tool management to operate effectively. For CAIR, completely removing the module yields 68.27%. Ablating Agent_{CI} alone still results in a 0.20% drop, showing that contextual integration is needed to reconcile verification signals; ablating Agent_{PV} causes a more pronounced decline to 69.53%, highlighting that independent perceptual verification is the more critical of the two agents. For DAAS, removing the tree-search mechanism entirely (w/o DAAS) yields 68.03%. Disabling difficulty awareness (69.87%) or the pruning mechanism (68.73%) also incurs clear penalties, demonstrating that both adaptive resource allocation and branch filtering are essential for high-quality search.

3) DAAS Efficiency Ablation: To assess the computational role of DAAS, Table VI reports average per-query results on one NVIDIA H200 GPU. Compared with UAV-MAS without DAAS, the full method introduces additional search computation while improving OA from 68.03% to 70.47%. More importantly, at a similar accuracy level, Majority Vote@3 applied to UAV-MAS without DAAS reaches 70.73% OA but requires 265.29 s, 36.39 MLLM calls, and 5.07 tool calls per query. Full UAV-MAS obtains 70.47% OA using 112.23 s, 25.60 MLLM calls, and 2.99 tool calls, reducing latency by 57.7% while requiring fewer model and tool calls. These results show that DAAS provides a better accuracy–cost trade-off than brute-force repeated sampling through targeted exploration.

TABLE VI
EFFICIENCY ABLATION OF DAAS ON ONE NVIDIA H200 GPU. MV3 DENOTES MAJORITY VOTE@3.

Method	OA (%)	Time (s)	MLLM Calls	Tool Calls
UAV-MAS w/o DAAS	68.03	88.43	12.13	1.69
UAV-MAS w/o DAAS + MV3	70.73	265.29	36.39	5.07
UAV-MAS	**70.47**	**112.23**	**25.60**	**2.99**

C. Visual Comparison

Figure 5 presents two representative cases where baselines fail and our UAV-MAS succeeds, collectively demonstrating all three modules. In both cases, DSPE's per-tool activation first selects a task-relevant tool subset rather than activating all available tools indiscriminately. In Figure 5(a), the extreme top-down viewpoint creates pose ambiguity: the shown open-source baselines misidentify cyclists as persons standing on the lawn. The DESCRIBE tool returns a high-confidence observation that conflicts with the prior belief; CAIR's step-level verification agent detects this conflict and updates the reasoning state accordingly, yielding the correct answer. This demonstrates that step-level verification is essential for correcting perceptual bias before it propagates through the reasoning chain. In Figure 5(b), the challenge is semantic distraction, manifesting in two distinct failure modes: the shown open-source baselines over-count to 8 by including parking-lot vehicles, while Gemini 3 Pro hallucinates a non-existent on-road vehicle and returns 4. DAAS assigns a low confidence score to the distractor-contaminated detection; two consecutive low scores then trigger branch pruning, and the search continues along alternative paths that correctly isolate the 3 on-road vehicles. This demonstrates that consecutive-confidence

TABLE VII
CROSS-DATASET GENERALIZATION ON THE CHOICE SUBSET. ALL VALUES ARE PERCENTAGES (%).

Method	OA	ILC	SII	CID	AttR	AssR	CSR
Qwen3-VL-8B Inst.	71.82	75.00	72.14	75.00	**65.00**	42.50	86.67
Qwen3-VL-8B Think.	69.09	70.00	68.57	**83.33**	47.50	47.50	78.33
UAV-MAS-8B	**75.23**	**85.00**	**77.14**	76.67	**65.00**	**50.00**	**91.67**

pruning effectively rejects misleading evidence and selects the globally more reliable reasoning trace.

D. Cross-Dataset Generalization on CHOICE

To evaluate cross-dataset generalization, we test UAV-MAS-8B on the independent CHOICE remote-sensing benchmark [57], using only generic image zooming and description tools, and compare it with the Qwen3-VL-8B Instruct and Thinking variants. We retain 440 questions whose answer formats are compatible with our evaluation and exclude Referring Expression Segmentation (RES) from Cross-instance Discernment because it requires pixel-level masks. Following the CHOICE taxonomy, ILC, SII, CID, AttR, AssR, and CSR denote Image-level Comprehension, Single-instance Identification, Cross-instance Discernment, Attribute Reasoning, Assessment Reasoning, and Common Sense Reasoning, respectively. As shown in Table VII, UAV-MAS-8B achieves the highest Overall Accuracy of 75.23%, outperforming the two baselines by 3.41 and 6.14 percentage points, respectively. These results demonstrate the effectiveness and generalizability of our method beyond the UAVQA-Bench distribution and UAV-specific perception tools.

VI. LIMITATIONS AND FUTURE WORK

Despite its encouraging performance, UAV-MAS has two main limitations. First, its iterative multi-agent reasoning and tool invocation introduce non-negligible inference latency. Although employing lightweight models and the DAAS strategy reduces unnecessary computation, the system remains substantially slower than a single forward pass. Consequently, it is currently more suitable for offline image analysis at a ground station than for real-time on-board inference. Second, its performance still has room for improvement. CAIR mitigates the propagation of local errors during iterative reasoning but cannot eliminate them entirely. In some cases, the model changes a correct answer to an incorrect one, fails to recognize critical visual information, or cannot objectively estimate the benefit of each reasoning step.

Future work will address these limitations from both efficiency and performance perspectives. For efficient deployment, we will investigate parallel agent execution, reusable visual-feature caching, more aggressive early-exit and dynamic-routing mechanisms, and model compression to reduce redundant computation and enable real-time on-board processing. To improve performance, we will explore stronger fine-grained visual perception, uncertainty-aware answer preservation and rollback mechanisms, and better-calibrated process-level verification for estimating step-wise reasoning gains. These directions are expected to make UAV-MAS faster, more reliable, and more practical for real-world UAV applications.

VII. CONCLUSION

In this paper, we introduce UAVQA-Bench, a comprehensive, fully human-annotated benchmark for UAV aerial image understanding and reasoning. It covers 6 capability dimensions across 16 tasks (in multiple-choice and visual grounding formats), with 1,500 samples drawn from 13 diverse public UAV datasets. We systematically evaluate a broad range of open-source and closed-source MLLMs as well as agent-based systems on UAVQA-Bench. Based on the results, we present UAV-MAS, a novel training-free multi-agent system designed to address the unique challenges of visual perception and reasoning in UAV scenarios. By integrating a Domain-Specific Perception Engine with advanced reasoning strategies—Context-Aware Iterative Refinement and Difficulty-Aware Adaptive Search—our system effectively overcomes the limitations of existing specialized models and general-purpose agents. Experiments demonstrate that UAV-MAS achieves state-of-the-art performance, enabling a 32B-parameter open-source model to surpass Gemini 3 Pro on UAVQA-Bench. These results underscore the potential of specialized multi-agent architectures in domain-specific applications. Future work will focus on optimizing real-time on-board inference and expanding the system's capabilities to dynamic video understanding tasks.

REFERENCES

[1] H. Zhang, K. Liu, Z. Gan, and G.-N. Zhu, "Uav-detr: efficient end-to-end object detection for unmanned aerial vehicle imagery," *arXiv preprint arXiv:2501.01855*, 2025.

[2] L. Madhuanand, F. Nex, and M. Y. Yang, "Self-supervised monocular depth estimation from oblique uav videos," *ISPRS journal of photogrammetry and remote sensing*, vol. 176, pp. 1–14, 2021.

[3] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European conference on computer vision*. Springer, 2024, pp. 216–233.

[4] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[5] X. Li, J. Ding, and M. Elhoseiny, "Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding," *Advances in Neural Information Processing Systems*, vol. 37, pp. 3229–3242, 2024.

[6] F. Wang, H. Wang, Z. Guo, D. Wang, Y. Wang, M. Chen, Q. Ma, L. Lan, W. Yang, J. Zhang *et al.*, "Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery?" in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 325–14 336.

[7] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[8] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.

[9] B. Zhao, J. Fang, Z. Dai, Z. Wang, J. Zha, W. Zhang, C. Gao, Y. Wang, J. Cui, X. Chen *et al.*, "Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces," *arXiv preprint arXiv:2503.06157*, 2025.

[10] S. Chakrabarty, "Yolo26: An analysis of nms-free end to end framework for real-time object detection," *arXiv preprint arXiv:2601.12882*, 2026.

[11] B. Gao, J. Tong, X. Chen, H. Yu, and Z. Li, "Dfir-detr: Frequency domain enhancement and dynamic feature aggregation for cross-scene small object detection," *arXiv preprint arXiv:2512.07078*, 2025.

[12] A. Khanpour, T. Wang, A. Vahidi-Shams, W. Ectors, F. Nakhaie, A. Taheri, and C. Claudel, "Uav-based intelligent traffic surveillance system: Real-time vehicle detection, classification, tracking, and behavioral analysis," *arXiv preprint arXiv:2509.04624*, 2025.

[13] Y. Chen, Z. Ye, H. Sun, T. Gong, S. Xiong, and X. Lu, "Global–local fusion with semantic information-guidance for accurate small object detection in UAV aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025, art. no. 4701115.

[14] Y. Xi, W. Jia, Q. Miao, J. Feng, J. Ren, and H. Luo, "Detection-driven exposure-correction network for nighttime drone-view object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024, art. no. 5605014.

[15] M. Maimaitijiang, H. Ghimire, S. Thapa, M. M. Billah, S. Sehgal, M. Singh, S. Kaushal, K. Poudel, S. Subedi, U. U. R. Janjua *et al.*, "Wheatai v1. 0: An ai-powered high throughput wheat phenotyping platform," *arXiv preprint arXiv:2601.08863*, 2026.

[16] D. E. Kharismawati and T. Kazic, "Maizestandcounting (masc): Automated and accurate maize stand counting from uav imagery using image processing and deep learning," *arXiv preprint arXiv:2510.07580*, 2025.

[17] O. Youme, J. M. Dembele, E. C. Ezin, and C. Cambier, "Panoptic segmentation of environmental uav images: Litter beach," *arXiv preprint arXiv:2508.15985*, 2025.

[18] P. Chen, T. Ouyang, K. Luo, W. Hong, and X. Chen, "Codrone: Autonomous drone navigation assisted by edge and cloud foundation models," *IEEE Internet of Things Journal*, 2025.

[19] Y. Lin, B. Xue, M. Zhang, S. Schofield, and R. Green, "Generalization evaluation of deep stereo matching methods for uav-based forestry applications," *arXiv preprint arXiv:2512.03427*, 2025.

[20] X. Shi, T. Tang, J. Chen, S. Lv, and Y. Liu, "Precise depth estimation by calculating affine transformation parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[21] S. Wang, K. Liu, J. Huang, and X. Li, "FLDet: Faster and lighter aerial object detector," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 5, pp. 4450–4463, 2025.

[22] F. Sun, D. Cheng, P. Zheng, T. Song, L. Chen, and Q. Kou, "TGCADNet: Text-guided context-aware detection via CLIP for small objects in UAV scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 6, pp. 7846–7859, 2026.

[23] Q. Jiang, J. Huo, X. Chen, Y. Xiong, Z. Zeng, Y. Chen, T. Ren, J. Yu, and L. Zhang, "Detect anything via next point prediction," *arXiv preprint arXiv:2510.12798*, 2025.

[24] Y. Cao, Z. He, L. Wang, W. Wang, Y. Yuan, D. Zhang, J. Zhang, P. Zhu, L. Van Gool, J. Han *et al.*, "Visdrone-det2021: The vision meets drone object detection challenge results," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 2847–2854.

[25] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[26] S. Wang, S. Li, Y. Zhang, S. Yu, S. Yuan, R. She, Q. Guo, J. Zheng, O. K. Howe, L. Chandra *et al.*, "Uavscenes: A multi-modal dataset for uavs," *arXiv preprint arXiv:2507.22412*, 2025.

[27] G. Li, J. Xu, Y. Zhao, and Y. Peng, "Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9098–9108.

[28] S. Zhao, H. Zhang, S. Lin, M. Li, Q. Wu, K. Zhang, and C. Wei, "Pyvision: Agentic vision with dynamic tooling," *arXiv preprint arXiv:2507.07998*, 2025.

[29] C. Scofield, "Multi-agent constraint factorization reveals latent invariant solution structure," *arXiv preprint arXiv:2601.15077*, 2026.

[30] Z. Ke, Y. Ming, A. Xu, R. Chin, X.-P. Nguyen, P. Jwalapuram, S. Yavuz, C. Xiong, and S. Joty, "Mas-orchestra: Understanding and improving multi-agent reasoning through holistic orchestration and controlled benchmarks," *arXiv preprint arXiv:2601.14652*, 2026.

[31] R. R. Rodriguez Jr, "Agent identity uri scheme: Topology-independent naming and capability-based discovery for multi-agent systems," *arXiv preprint arXiv:2601.14567*, 2026.

[32] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[33] Q. Ma, C. Guo, Z. Tian, S. Wang, J. Xiao, Y. Yue, and Z. Zhang, "Paper2rebuttal: A multi-agent framework for transparent author response assistance," *arXiv preprint arXiv:2601.14171*, 2026.

[34] A. Adimulam, R. Gupta, and S. Kumar, "The orchestration of multi-agent systems: Architectures, protocols, and enterprise adoption," *arXiv preprint arXiv:2601.13671*, 2026.

[35] H. Lee, "Motion-to-response content generation via multi-agent ai system with real-time safety verification," *arXiv preprint arXiv:2601.13589*, 2026.

[36] S. Hui, D. Yanfeng, H. Ma, C. Xu, K. Jin, L. Zu, C. Zhong, G. Wang, W. Cai *et al.*, "Agentgc: Evolutionary learning-based lossless compression for genomics data with llm-driven multiple agent," *arXiv preprint arXiv:2601.13559*, 2026.

[37] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The eleventh international conference on learning representations*, 2022.

[38] I. Bozcan and E. 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)*. IEEE, 2020, pp. 8504–8510.

[39] C. Zhang, G. Huang, L. Liu, S. Huang, Y. Yang, X. Wan, S. Ge, and D. Tao, "Webuav-3m: A benchmark for unveiling the power of million-scale deep uav tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9186–9205, 2022.

[40] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.

[41] U. Graz, "Semantic drone dataset," 2019. [Online]. Available: http://dronedataset.icg.tugraz.at/

[42] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022.

[43] W. Cai, K. Jin, J. Hou, C. Guo, L. Wu, and W. Yang, "Vdd: Varied drone dataset for semantic segmentation," *Journal of Visual Communication and Image Representation*, vol. 109, p. 104429, 2025.

[44] Y. Chen, Y. Wang, P. Lu, Y. Chen, and G. Wang, "Large-scale structure from motion with semantic constraints of aerial images," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 347–359.

[45] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.

[46] H. Florea, V.-C. Miclea, and S. Nedevschi, "Wilduav: Monocular uav dataset for depth estimation tasks," in *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2021, pp. 291–298.

[47] C. Feng, Z. Chen, X. Li, C. Wang, J. Yang, M.-M. Cheng, Y. Dai, and Q. Fu, "Hazydet: Open-source benchmark for drone-view object detection with depth-cues in hazy scenes," *arXiv preprint arXiv:2409.19833*, 2024.

[48] P. Zhu, T. Peng, D. Du, H. Yu, L. Zhang, and Q. Hu, "Graph regularized flow attention network for video animal counting from drones," *IEEE Transactions on Image Processing*, vol. 30, pp. 5339–5351, 2021.

[49] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *European conference on computer vision*, vol. 7, 2016.

[50] H. Lin, S. Chen, J. H. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," *arXiv preprint arXiv:2511.10647*, 2025.

[51] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[52] V. Team, W. Hong, W. Yu, X. Gu, G. Wang, G. Gan, H. Tang, J. Cheng, J. Qi, J. Ji, L. Pan, S. Duan, W. Wang, Y. Wang, Y. Cheng, Z. He, Z. Su, Z. Yang, Z. Pan, A. Zeng, B. Wang, B. Chen, B. Shi, C. Pang, C. Zhang, D. Yin, F. Yang, G. Chen, J. Xu, J. Zhu, J. Chen, J. Chen, J. Chen, J. Lin, J. Wang, J. Chen, L. Lei, L. Gong, L. Pan, M. Liu, M. Xu, M. Zhang, Q. Zheng, S. Yang, S. Zhong, S. Huang, S. Zhao, S. Xue, S. Tu, S. Meng, T. Zhang, T. Luo, T. Hao, T. Tong,

W. Li, W. Jia, X. Liu, X. Zhang, X. Lyu, X. Fan, X. Huang, Y. Wang, Y. Xue, Y. Wang, Y. Wang, Y. An, Y. Du, Y. Shi, Y. Huang, Y. Niu, Y. Wang, Y. Yue, Y. Li, Y. Zhang, Y. Wang, Y. Wang, Y. Zhang, Z. Xue, Z. Hou, Z. Du, Z. Wang, P. Zhang, D. Liu, B. Xu, J. Li, M. Huang, Y. Dong, and J. Tang, "Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2507.01006

[53] Q. Team, "Qwen3.5: Accelerating productivity with native multimodal agents," February 2026. [Online]. Available: https://qwen.ai/blog?id=qwen3.5

[54] OpenAI, "Introducing gpt-5.2," https://openai.com/index/introducing-gpt-5-2/, 2025, accessed: 2026-01-28.

[55] Google DeepMind, "Gemini 3 pro and gemini 3 flash models," https://ai.google.dev/, 2025, accessed: 2026-01-15.

[56] QwenLM, "Qwen-agent: An agent framework based on qwen," https://github.com/QwenLM/Qwen-Agent, 2024.

[57] X. An, J. Sun, Z. Gui, and W. He, "Choice: Benchmarking the remote sensing capabilities of large vision-language models," in *Advances in Neural Information Processing Systems*, vol. 38, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2025/hash/befe25a01cf4dbe9635e85f835d31250-Abstract-Datasets_and_Benchmarks_Track.html

Haoyu Zhang received the B.S. degree in electronic science and technology from the School of Information Science and Technology, Fudan University, Shanghai, China, in 2025. He is currently pursuing the Ph.D. degree with the College of Future Information Technology, Fudan University. His research interests include computer vision, multimodal large models, and intelligent agent systems.

Shuoxun Zhang received the B.S. degree in Communication Engineering from Shanghai University, Shanghai, China, in 2023. He is currently pursuing a master's degree at the College of Future Information Technology, Fudan University. His main research interests include multimodal large models and intelligent agent systems.

Peng Ye is currently a postdoctoral fellow at MMLab of The Chinese University of Hong Kong and a scientific research advisor of Shanghai AI Laboratory. He received the Ph.D. degree from Fudan University, Shanghai, China. His research interests include autonomous agents, (M)LLMs, foundation models, and efficient design and optimization. He has published papers in leading journals and conferences, including IEEE TPAMI, IJCV, CVPR, ICCV, ECCV, NeurIPS, ICML, ACM MM, and ICME. He also serves as a reviewer or program committee member for journals and conferences including IEEE TPAMI, IJCV, CVPR, ECCV, ICCV, and NeurIPS.

Lin Zhang received the B.S. degree in electronic engineering from Fudan University, Shanghai, China, in 2022, where he is currently pursuing the Ph.D. degree with the College of Future Information Technology. His main research interests include computer vision and transfer learning.

Jiakang Yuan is currently pursuing the Ph.D. degree in electronic engineering with the College of Future Information Technology, Fudan University. He received the bachelor's degree in electronic engineering from Fudan University in 2022. His research interests include multimodal reasoning, multi-agent systems, and spatial intelligence. He has published papers in leading journals and conferences, including CVPR, ICCV, ECCV, NeurIPS, and IEEE TPAMI, and serves as a reviewer for journals and conferences including IEEE TIP, IEEE TCSVT, CVPR, ECCV, and ICCV.

Shenghong Yi received the B.S. degree in Intelligent Science and Technology from Fudan University, Shanghai, China, in 2026. He is currently working toward the Ph.D. degree at the College of Future Information Technology. His research interests include computer vision and multimodal large models.

Yuening Wang is currently pursuing the bachelor's degree in Electronic Information Science and Technology at the College of Future Information Innovation, Fudan University, Shanghai, China, and is expected to graduate in June 2027.

Tao Chen (Senior Member, IEEE) received the Ph.D. degree in information engineering from Nanyang Technological University, Singapore, in 2013. He was a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore, from 2013 to 2017, and a Senior Scientist with the Huawei Singapore Research Center from 2017 to 2018. He is currently a Professor with the College of Future Information Technology, Fudan University, Shanghai, China. His main research interests include efficient computer vision, multimodal visual analysis, large VLM compression, and their applications in embodied robot intelligence, scene understanding, and reconstruction. He has published over 200 papers in international journals and conferences such as IEEE TPAMI, IEEE TIP, IJCV, CVPR, and NeurIPS. He has served in area chair and senior program committee roles for conferences such as AAAI, ICLR, and PRCV. He received the IJCAI 2025 Distinguished Paper Award.

Supplementary Material for: *Advancing MLLM-based UAV Image Understanding and Reasoning: A Benchmark and a Training-Free Multi-Agent System*

APPENDIX A
MORE DETAILS ABOUT UAVQA-BENCH

A. Question Templates

To ensure the linguistic diversity of UAVQA-Bench, we develop a variety of question templates for each task. This section presents representative templates that define the interaction logic for the agent.

1) Existence Detection (ED):

- **Scene Presence**
  - Is any {object} present here?
  - Is any {object} visible in the image?
  - Is there any {object} in the scene?
  - Does this image contain any {object}?
  - Can you see any {object} in this picture?
- **Conditional Presence**
  - Is any {object} present {condition}?
  - Is a/an {object} {condition} in the image?
  - Is there any {object} {condition} in this image?
  - Does the image contain a/an {object} {condition}?
  - Can you see any {object} {condition} in this picture?
- **Multi-Target Presence**
  - Which of the listed objects can be seen in the image?
  - Does the image show any of the objects listed below?
  - Which of the following objects are present in the image?
  - From the options below, which objects appear in the picture?
  - Among the following choices, which ones are present in the photo?

2) Category Recognition (CR):

- **Regional Classification**
  - Identify the object found in $\{<x_1><y_1><x_2><y_2>\}$.
  - Classify the object in region $\{<x_1><y_1><x_2><y_2>\}$.
  - Which class does the object in $\{<x_1><y_1><x_2><y_2>\}$ belong to?
  - What is the category of the object located at $\{<x_1><y_1><x_2><y_2>\}$?
  - Label the region $\{<x_1><y_1><x_2><y_2>\}$ with the correct object category.

3) Quantity Awareness (QA):

- **Scene Counting**
  - Provide the quantity of {object}.
  - Count all the {object} in this picture.
  - How many {object} are in this image?
  - How many instances of {object} can you see?
  - Can you count the number of {object} present?
- **Regional Counting**
  - Provide the quantity of {object} {condition}.
  - Count all the {object} {condition} in this picture.
  - What is the total number of {object} visible {condition}?
  - How many {object} appear inside this bounding box $\{<x_1><y_1><x_2><y_2>\}$?
  - What is the total number of {object} in the bounding box $\{<x_1><y_1><x_2><y_2>\}$?
  - Can you count the number of {object} within the bounding box $\{<x_1><y_1><x_2><y_2>\}$?
- **Number Comparison**

- Do {object}$_A$ outnumber {object}$_B$?
- Do {object}$_A$ exceed {object}$_B$ in quantity?
- Are {object}$_A$ less abundant than {object}$_B$?
- Are {object}$_A$ more numerous than {object}$_B$?
- Is it true that {object}$_A$ is fewer than {object}$_B$?

4) Fine-grained Attribute Perception (FAP):

- **Regional Attribute Recognition**
  - What color is the {object} in {<x_1><y_1><x_2><y_2>}?
  - Identify the color of the {object} in {<x_1><y_1><x_2><y_2>}.
  - Which color does the {object} in {<x_1><y_1><x_2><y_2>} have?
  - Identify the driving direction (relative to the camera) of the {object} in {<x_1><y_1><x_2><y_2>}.
  - What is the driving direction (relative to the camera) of the {object} located at {<x_1><y_1><x_2><y_2>}?
- **Function Recognition**
  - What is the key function of the {object} bounded by {<x_1><y_1><x_2><y_2>}?
  - Which functional category does the {object} in {<x_1><y_1><x_2><y_2>} belong to?
  - The {object} within the bounding box {<x_1><y_1><x_2><y_2>} falls under which functional category?
- **Safety Landing**
  - Identify the safest landing area for a drone among the given options.
  - Among the options below, where is the best place for a drone to land?
  - Select the optimal landing spot for a drone from the following choices.
  - Which location is most ideal for a drone landing in the options provided?
  - From the following, choose the most appropriate site for a drone to land.

5) Spatial Relationship Understanding (SRU):

- **Spatial Relation**
  - Relative to the object in {<x_1><y_1><x_2><y_2>}$_A$, where is the object in {<x_1><y_1><x_2><y_2>}$_B$?
  - As seen by the object in {<x_1><y_1><x_2><y_2>}$_A$, in what direction is the object in {<x_1><y_1><x_2><y_2>}$_B$?
  - From the viewpoint of the object in {<x_1><y_1><x_2><y_2>}$_A$, what is the direction to the object in {<x_1><y_1><x_2><y_2>}$_B$?
  - Assuming the UAV's forward direction corresponds to the top of the image, at which clock position is the object in {<x_1><y_1><x_2><y_2>} located?
  - Given that the top of the image is treated as the front of the UAV, what is the clock position of the object within {<x_1><y_1><x_2><y_2>} relative to the UAV?
- **Height Comparison**
  - Which between {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$ is determined to be at a higher position?
  - Between the two positions {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$, which one is judged to be higher in the frame?
  - Please judge which object is higher in the frame between {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$?
  - Can you judge which contains an object at a higher position between {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$?
  - Please evaluate the two positions {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$, and indicate which one is higher in the vertical direction.
- **Distance Comparison**
  - Which between {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$ is measured to be closer to the shooting point?
  - Between {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$, which object has a smaller depth value, meaning it is closer?
  - Please evaluate the two positions {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$, and indicate which one is closer to the drone.
  - Please judge which object is closer to the drone between {<x_1><y_1><x_2><y_2>}$_A$ and {<x_1><y_1><x_2><y_2>}$_B$?

6) Visual Grounding (VG):

- **Simple Object Grounding**
  - Locate the {object}.
  - Point out the {object}.
  - Could you point out the {object}?
  - Draw the position of a/an {object}.

- Can you show me where the {object} is?

- **Complex Semantic Grounding**
  - Please give me the location of {semantics}.
  - Please indicate the position of {semantics}.
  - Can you point out the location of {semantics}?
  - Could you tell me the location for {semantics}?
  - In the current image, where can I find {semantics}?
- **Highest Object Grounding**
  - Find the area with the tallest building.
  - Which area in the image has the tallest building?
  - Judge which building area in the image has the highest floors.

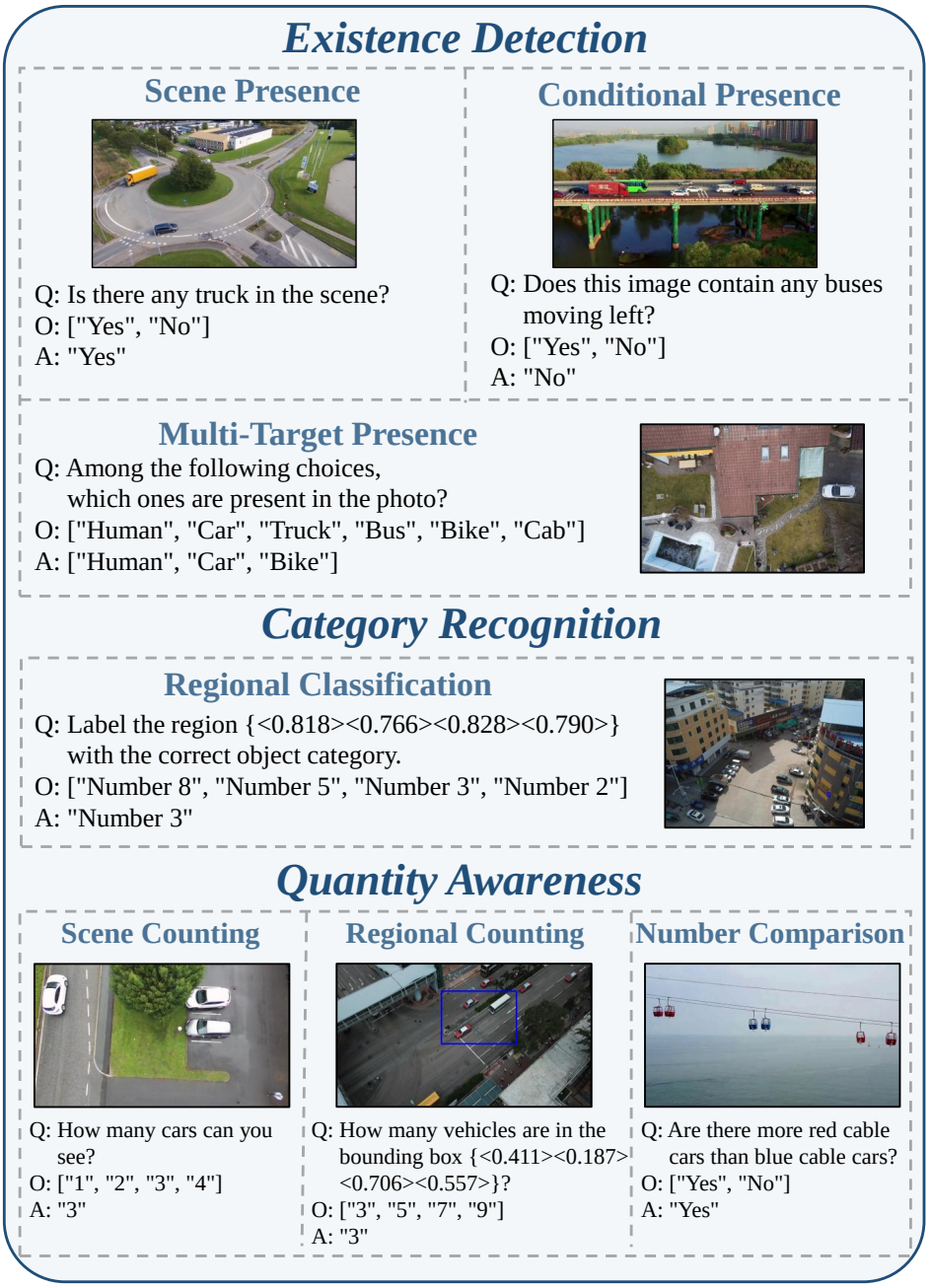


Fig. 1. Sample instances for tasks in Existence Detection, Category Recognition, and Quantity Awareness.

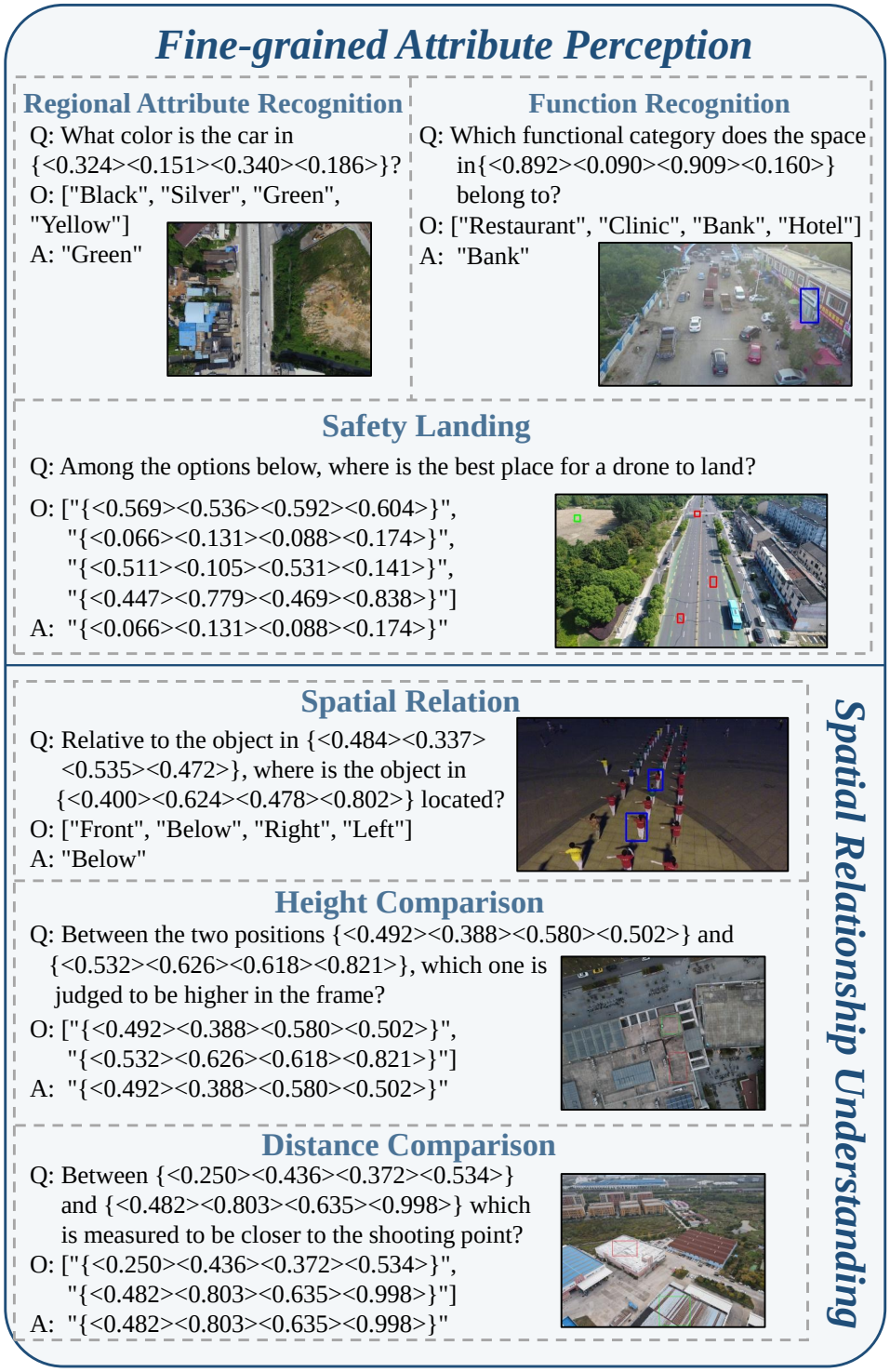


Fig. 2. Sample instances for tasks in Fine-grained Attribute Perception and Spatial Relationship Understanding.

7) Template Visualizations: In this section, we present additional examples from UAVQA-Bench. Figure 1 provides sample instances and visual prompts for the tasks of Existence Detection, Category Recognition, and Quantity Awareness. Figure 2 illustrates examples for Fine-grained Attribute Perception and Spatial Relationship Understanding. Finally, Figure 3 showcases samples for the Visual Grounding tasks. In these samples, “Q”, “O”, and “A” denote the Question, Options, and Ground-truth Answer, respectively. For clarity, the sample figures include bounding boxes in different colors: red boxes indicate incorrect candidate answers, green boxes denote the correct answer, and blue boxes mark the object or region referred to in the question. These bounding boxes are shown only for visualization in the paper and are not present in the actual images provided to the model during inference.

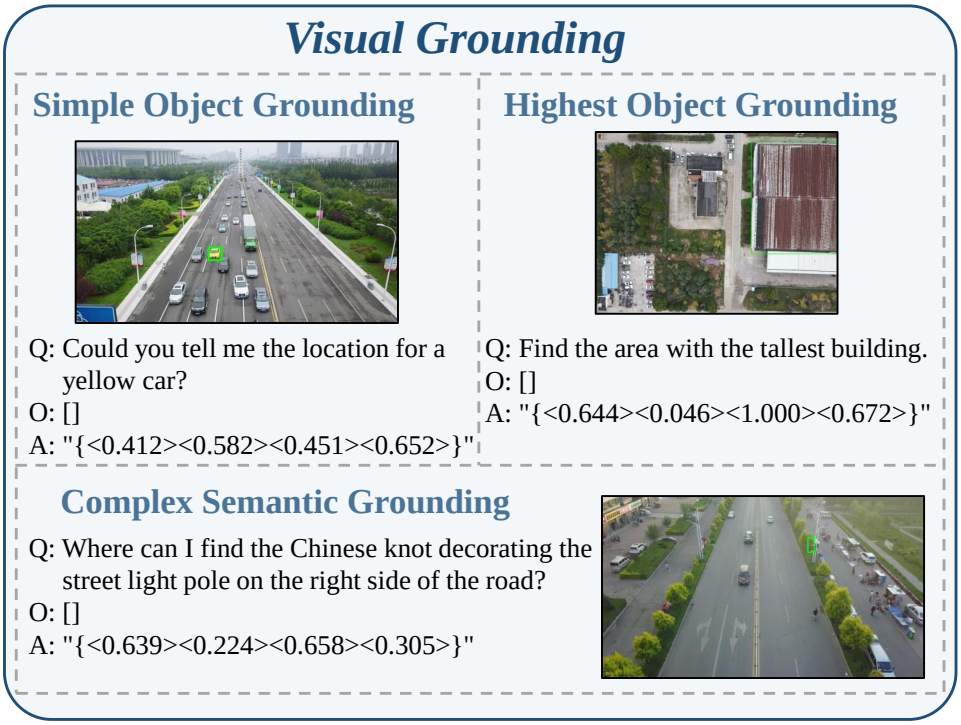


Fig. 3. Sample instances for tasks in Visual Grounding.

APPENDIX B
MORE DETAILS OF OUR METHOD

In this section, we provide the implementation settings and agent configurations used in our experiments.

A. Implementation Details

All MLLM calls use a temperature of 0.7 and a maximum of 8,192 new tokens. For DAAS, the maximum depth is $D = 5$, and the search width is set adaptively to $W = \min(3, |\mathcal{T}_{\text{opt}}|)$, where $|\mathcal{T}_{\text{opt}}|$ is the number of tools selected by DSPE. Direct MLLM baselines use their default system prompts with the same image–query input structure, temperature, and token budget. All experiments are conducted on NVIDIA H200 GPUs.

B. Agent Configurations

We detail the configuration of each agent in our framework in Table I. We utilize the Qwen3-VL series as the backbone MLLM. Specifically, we employ the *Instruct* variant for agents that require precise execution of specialized instructions (e.g., Agent_{TS}, Agent_{SR}), and the *Thinking* variant for the Long-Chain Answer Agent (Agent_{LA}) to leverage its enhanced capabilities in complex reasoning and synthesis. It is worth noting that the *Instruct* and *Thinking* models share the identical architectural design, differing only in their parameter weights which are tuned for instruction compliance and deep reasoning, respectively.

TABLE I
AGENT CONFIGURATIONS AND MODEL TYPES USED IN UAV-MAS.

Agent Role	**Model Type**
Tool Select Agent (Agent_{TS})	Qwen3-VL Instruct
Strategic Reasoning Agent (Agent_{SR})	Qwen3-VL Instruct
Perceptual Verification Agent (Agent_{PV})	Qwen3-VL Instruct
Contextual Integration Agent (Agent_{CI})	Qwen3-VL Instruct
Long-Chain Answer Agent (Agent_{LA})	Qwen3-VL Thinking
Score Agent (Agent_{SC})	Qwen3-VL Instruct