

# Supplementary Material of AirCopBench: A Benchmark for Embodied Collaborative Perception and Reasoning

## A Related Work

### A.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) (Zhang et al. 2024a) extend traditional language models by integrating multiple data modalities, such as text, images, and video, enabling more complex visual reasoning and understanding across diverse data types. Models like BLIP (Li et al. 2022), GPT-4V (Yao et al. 2024), and Gemini (Comanici et al. 2025) demonstrate the power of combining vision and language for tasks such as image classification and captioning. Similarly, VisualBERT (Li et al. 2019), Flamingo (Alayrac et al. 2022), and LLaVA (Liu et al. 2024) focus on vision-language fusion for VQA tasks, allowing reasoning over both images and text. Recent work has explored MLLM-based question answering for multi-view perception, but challenges remain in role-based reasoning and collaboration among agents, particularly in dynamic, real-world scenarios with perception degradation.

### A.2 Collaborative Perception Benchmarks

Recent research has introduced several collaborative perception benchmarks for multi-UAV systems. Coperception-UAV (Hu et al. 2022) focus on collaborative object detection and semantic segmentation, while MDMT (Liu et al. 2023) evaluates multi-task perception in dynamic environments. UAV3D (Sunderraman, Ji et al. 2024) and AeroCollab3D (Tian et al. 2024) extend 3D perception in collaborative settings with more scenarios, and Air-Co-Pred (Wang et al. 2024c) targets occlusion-aware prediction. For VQA-based benchmarks on multi-view images, UrBench (Zhou et al. 2025) and MUIRBENCH (Wang et al. 2024a) focus on urban and multi-image environments, respectively, while All-Angles Bench (Yeh et al. 2025) targets multi-angle perception. Although these benchmarks advance collaborative perception, they focus on specific modalities or perception tasks, limiting applicability to complex scenarios. In contrast, our AirCopBench addresses a wider range of challenges, including various perception degradation, multimodal inputs, and collaborative decision-making, offering a more comprehensive evaluation framework for multi-UAV

systems. This makes AirCopBench a valuable tool for advancing aerial collaborative perception. The comparison of AirCopBench with the existing benchmarks is in Tab.1.

## B Dataset Examples

To better illustrate the proposed dataset, we present VQA examples along with corresponding multi-view images captured by multiple UAVs in the real world (Liu et al. 2023), the EmbodiedCity simulator (Gao et al. 2024), the Coperception-UAV dataset (Hu et al. 2022), and the AeroCollab3D dataset (Tian et al. 2024), covering all 14 task types across 4 dimensions. The dataset examples from the 4 data sources are shown in Fig. 1, Fig. 2, Fig. 3, and Fig. 4, respectively. The complete dataset can be found in [https://drive.google.com/drive/folders/1MeCM2\\_MA5A-1XsIgvSZZacCWk-sistgp?usp=sharing](https://drive.google.com/drive/folders/1MeCM2_MA5A-1XsIgvSZZacCWk-sistgp?usp=sharing).

## C Benchmark Generation Pipeline

### C.1 Details of Data Collection

**Selected Simulator Advantages.** The choice of simulators is based on several key advantages: a) Realistic and High-Fidelity Environments: All three simulators provide highly realistic environments that closely replicate real-world conditions, ensuring practical data relevance. b) Support for Multi-Agent Collaboration: These simulators enable the simulation of multiple agents (e.g., UAVs, robots) interacting within a shared space, making them ideal for studying collaborative perception tasks and information exchange. c) High-Degree-of-Freedom Construction of Perception Degradation: They allow for a variety of environmental setups, creating scenarios with perception degradation factors like occlusions and interference, essential for testing system robustness. d) Scalability and Flexibility: The simulators offer scalability in agent numbers, task complexity, and environmental variables, allowing flexible adjustments to suit different refinement needs.

**Representative Scenarios Constructions.** In selecting image pairs and constructing representative scenarios, we take into account a wide range of challenging perception degradation types that can affect the quality of sensor data. These include factors such as obstacle occlusion, shadow coverage, motion blur, low signal-to-noise ratio (SNR), data



Figure 1: Examples of real-world images captured by real drones and their corresponding VQA pairs.

loss, lighting imbalance, complex backgrounds, objects outside the field of view (FoV), small target sizes, and varying weather conditions. Each of these degradation types represents a unique challenge to perception systems, making it critical to incorporate them into the scenario design for testing robustness under realistic conditions. Additional examples of these scenarios can be found in Fig. 1-4.

**Flexible Settings.** In the context of multi-UAV collaborative perception for data collection, we develop custom scripts that provide highly flexible settings to simulate a wide range of scenarios in EmbodiedCity. This flexibility is essential for testing UAV coordination, perception systems, and mission strategies under diverse conditions. a) Drone Setup: Our data collection scripts allow users to change the number of UAVs, both targets and observers, in the simulation, helping model different observation group configurations with varying levels of complexity. b) Initial Position and Postures of UAVs: We adjust the initial positions, altitudes, and orientations of UAVs during data collection. This customization simulates different starting conditions, such as UAVs positioned at varying distances or angles, which is crucial for evaluating perception systems in diverse spa-

tial setups and real-world scenarios. c) Camera Settings of Observer UAVs: We modify the camera settings of observer UAVs to use “lookdown,” “horizon,” or “oblique” perspectives. This allows for testing how different viewing angles impact tracking and recognition tasks. Each camera setting provides a unique perspective, which is vital for assessing the performance of perception algorithms in various observation scenarios. d) Motion Trajectories of UAVs: This setting enables customized drone trajectories of the target, allowing us to simulate different movement behaviors, from simple circular motion to more complex, dynamic paths.

## C.2 Details of Data Annotation

In this work, we introduce both object- and event-level labeling to generate rich and rational annotations tailored for multi-drone collaborative perception with challenging perception degradation. Despite the original object-level annotations, we also design a human annotation interface using Label-Studio (Tkachenko et al. 2020-2025) for event-level labeling on specific collaborative events, as shown in Fig. 20. The event-labeling information encompasses image quality scoring, perception usability assessments, reasoning for perception degradation, and aspects of collaborative de-



Figure 2: Examples of simulated images collected from EmbodiedCity and their corresponding VQA pairs.

cision analysis, including when and why inter-UAV communication is required, who is suited for information retrieval, and what observational data should be shared.

### C.3 Details of Question Generation

To produce high-quality VQA pairs, we employ three approaches, model-based, rule-based, and human-based generation, ensuring the validity of our dataset.

**Model-based Generation.** This section uses LLMs to generate questions for task types requiring high diversity and contextual richness. Specifically, we apply four prompting strategies to guide the model in generating more relevant and coherent questions. The prompting examples and templates for model-based VQA generation are shown in E.4.

- **Task Decomposition.** We begin by decomposing the overall VQA generation task, which encompasses all 14 collaborative perception tasks, into sub-tasks based on the independent complexity, requirements, and cognitive abilities of each task type. Instead of using a single VQA generation prompt for all tasks, we create specific functions for each task, call them sequentially, and merge their results into the final data structure. This approach

enhances the model’s success rates, simplifies debugging and retries, and improves the quality of generated questions by allowing the model to focus on each task individually.

- **Role-playing.** To improve task efficiency and reduce redundancy, we suggest incorporating global rules into the system role’s content. This allows the user prompt to focus solely on the specific task at hand. For example, by specifying the system’s role as an expert assistant and setting rules like “output must be in JSON,” these constraints are applied globally, leaving the user prompt simpler and more concise. This approach not only separates the responsibilities of the system and user prompts but also reduces token consumption by eliminating repetitive instructions in every user request.
- **CoT Prompting.** For tasks requiring deep visual understanding, such as Scene Description, Scene Comparison, and Causal Assessment, we adopt a two-step CoT Prompting strategy that combines model pre-processing and generation. First, the model generates an intermediate understanding of the images, such as textual descriptions or captions for each UAV’s perspective. In the sec-





Figure 3: Examples of simulated images collected from Coperception-UAV and their corresponding VQA pairs.

ond step, on the basis of these descriptions, the model generates the final question and answer pair that better fits the image content and question type. This approach simplifies the complex “image-to-question” task into two more manageable steps: “image-to-text” and “text-to-question.” By generating captions or reasoning first, followed by question creation, this method improves both task manageability and question quality.

- **Few-shot Learning.** For tasks requiring complex reasoning, such as Object Grounding or Why to Collaborate, incorporating a few-shot learning approach in the prompt can significantly improve model performance. By providing a structured example in a few-shot format, the model learns to generate responses in the desired JSON format and question style. This approach enforces stricter format constraints and guides the model to focus on relevant details, ensuring more accurate and contextually ap-

propriate questions.

**Rule-based Generation.** For tasks where answers can be directly derived from annotated data, such as Object Recognition, Object Counting, Quality Assessment, and When to Collaborate, rule-based generation is the most efficient approach. This method eliminates the need for complex visual reasoning and significantly reduces API call costs, ensuring complete accuracy and control over the generated results. For example, in Object Recognition and Counting tasks, annotations can directly provide accurate object counts or unique labels, which are then used to generate questions and answer options based on predefined templates. This approach streamlines the question generation process, ensuring consistency and precision across tasks. The rule-based prompting templates for VQA generation are shown in E.4.

**Human-based Generation.** This generation method plays a critical role in tasks such as Observing Posture

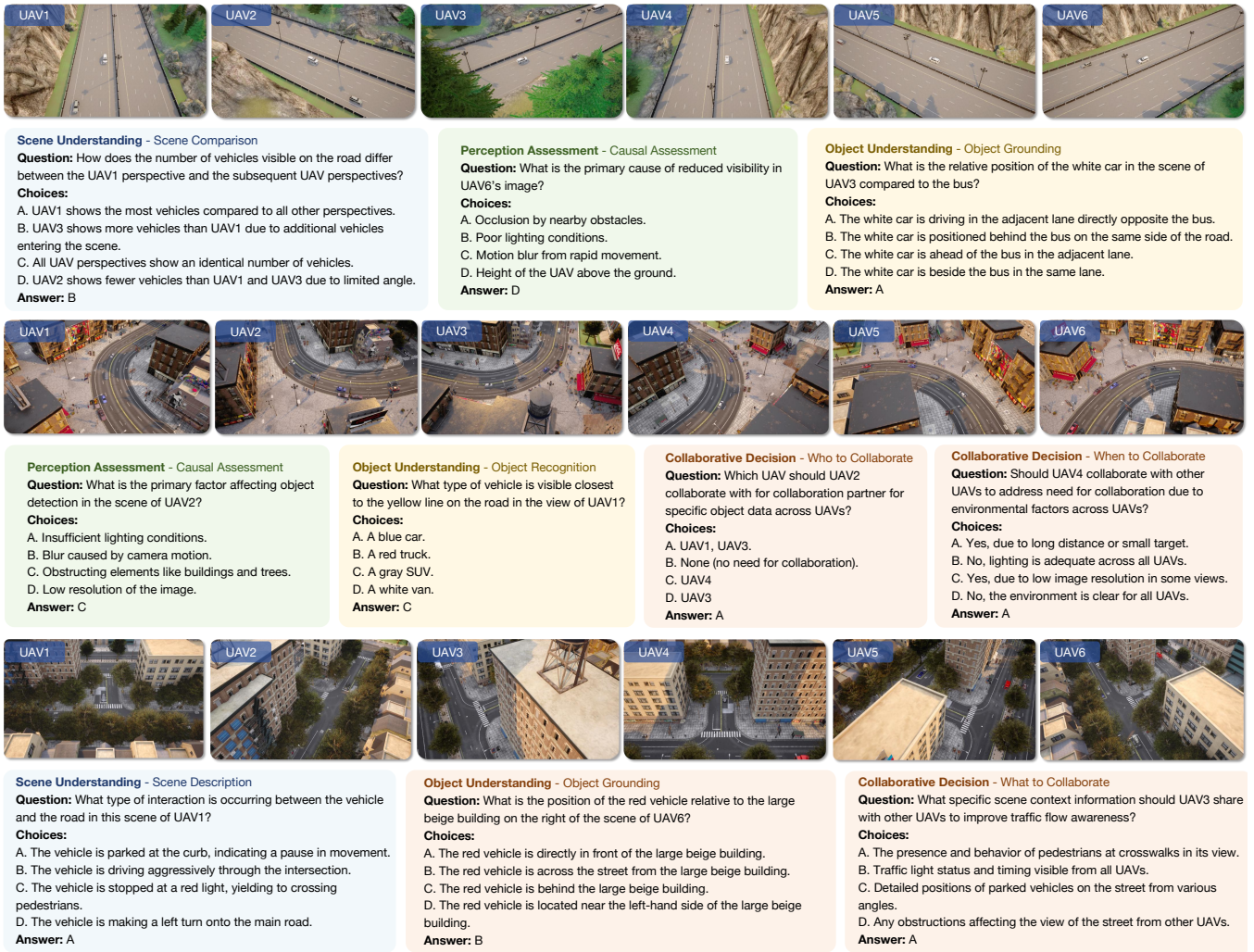


Figure 4: Examples of simulated images collected from AeroCollab3D and their corresponding VQA pairs.

and Object Matching, which require advanced multi-image reasoning and a deep spatial understanding—skills that current models typically struggle to achieve. By integrating human experts into the annotation process, we ensure that the generated questions are able to capture intricate visual details, the complexities of real-world scenarios, and the nuanced logical relationships that may arise. The expertise of these human annotators allows for the development of high-quality, contextually rich questions that mirror the challenges found in real-world situations, while also aligning with human reasoning processes. Their involvement significantly enhances the quality and relevance of the annotations, making them more applicable to practical applications. Several human-generated VQA examples are given in E.4.

**Hybrid Generation.** This phase also involves hybrid question generation. When model-based generation fails, rule-based generation fills the gaps, or model-based generation follows rule-based generation to enhance question va-

riety and complexity. Furthermore, human can assist in improving machine-generated questions, enhancing efficiency and relevance.

#### C.4 Details of Quality Control

Table 1: Comparison of MATTR scores for window size 50 between our dataset and llava-instruct-150k.

Dataset	MATTR (window $w = 50$ )
Our Dataset	0.7371
llava-instruct-150k	0.7834

**Standard Examination.** This measure evaluates the quality of generated VQA pairs by examining four key criteria that determine their completeness and usefulness: a) Required Content ensures that all necessary information is included within the VQA pair. This criterion assesses whether the pair contains all the essential components needed to fully

address the question. Any VQA pair that lacks critical information or is incomplete will be flagged for revision, ensuring that the final output provides all the relevant details; b) Format Consistency is concerned with maintaining a uniform structure, wording, and presentation throughout the VQA pairs. This criterion checks if the question and answer formats are consistent across different pairs, ensuring that each pair adheres to a standardized template. Pairs that deviate from this uniformity will be identified for correction; c) Answer Validity assesses the correctness and relevance of the provided answer options. This includes verifying that all answer choices are accurate and that the correct answer is included among them. Any VQA pair that features incorrect or irrelevant options will be filtered out, ensuring that only valid pairs remain; d) Question Length evaluates whether the questions posed are sufficiently detailed to avoid ambiguity. A question that is too vague or lacks clarity can lead to confusion and imprecise answers. This criterion ensures that questions are comprehensive and clear, offering enough context to guide the response accurately.

For each of the four criteria, if a VQA pair meets the required standard, it will earn one point. The final score for each VQA pair is determined by the total number of criteria satisfied, with a maximum possible score of 4 points. Pairs that score below 4 will be deleted, while those that meet all four conditions will remain in the final dataset.

**Blind Filtering.** This measure is designed to eliminate questions that can be answered purely through common sense, without the need for specific visual input. By leveraging multiple MLLMs (3 MLLMs in this work), the system predicts the answers to the questions without incorporating the multi-view image input. If all MLLMs provide correct answers, the question is removed from the dataset. This outcome suggests that the question can be accurately answered based on general perceptual knowledge alone, and does not require the detailed visual understanding of the multi-view images. This approach ensures that only those questions necessitating complex image-based reasoning and multi-view analysis are retained, thereby enhancing the dataset’s focus on challenging perceptual tasks. (Zhao et al. 2025). Conversely, if all MLLMs answer incorrectly, the question is also filtered out, as this suggests the question may be inherently flawed, ambiguous, or beyond the reasonable scope of the task. The dual filtering criterion guarantees that the final dataset includes only questions necessitating authentic multi-view visual reasoning, characterized by being correctly answered by some but not all MLLMs in the absence of visual input.

**Human Refinement.** To further address the VQA pairs that still have following issues:

- **Ambiguous Questions** usually lack clear definitions, leading to confusion in task understanding. Poorly framed questions can either be too broad or too narrow, hindering the ability of UAVs to collaborate effectively. Such questions fail to capture the task’s complexity, influencing the evaluation accuracy of model’s capabilities. Clear, well-defined questions are essential for effective training and evaluation in collaborative perception tasks. Several

refinement examples of ambiguous questions are given in Tab. 2.

- **Invalid Options** with no correct answer, duplicates, or those irrelevant to the question disrupt the decision-making process. Additionally, when options lack clear distinction, agents in the system struggle to make accurate choices. For effective multi-agent collaboration, it is crucial to ensure that the options are relevant, distinct, and properly aligned with the question. Several refinement examples of invalid options are given in Tab. 2.
- **Incorrect Answers** include choosing the wrong answer, missing the correct one, or selecting multiple correct answers. Such errors hinder agents’ ability to collaborate effectively, making it crucial to ensure accurate and consistent answer selection for task success. Several refinement examples of incorrect answers are given in Tab. 2.

**Lexical Diversity Comparison.** To quantify the lexical richness of our dataset relative to a high-quality reference, we compute the Moving-Average Type-Token Ratio (MATTR) (?) with a window size of 50 tokens shown in Tab. 1. MATTR is defined as:

$$\text{MATTR}_w = \frac{1}{N - w + 1} \sum_{i=1}^{N-w+1} \frac{|\{t_i, t_{i+1}, \dots, t_{i+w-1}\}|}{w},$$

where  $N$  is the total number of tokens in the corpus,  $w$  is the window size, and  $\{t_i, \dots, t_{i+w-1}\}$  is the set of unique tokens in the  $i$ -th window.

## D Benchmark Statistics

The training set and the test set contain approximately 13.6k and 1k questions, respectively. We present wordcloud visualizations of the textual content from the real-world dataset and three simulator datasets and their corresponding real-world scenarios, as illustrated in Figure 5.

## E Experimental Details

### E.1 Computation Details

All local model inference and fine-tuning is performed on  $4 \times$  A800-SXM4-80GB. The full code of our program is available at <https://github.com/zhajirong/AirCopBench>.

### E.2 Baseline Introduction

Our evaluation covers both proprietary and open-source MLLMs trained to receive multi-image inputs. The evaluated models as well as random and human baseline are briefly introduced as follows:

**Random.** A random baseline model serving as the lowest performance benchmark for comparison.

**Human.** Human expert performance baseline representing the upper limit of human-level capability on this task.

**Step-1o-turbo.** A reasoning-optimized turbo version model from stepfun that demonstrates excellent performance across multiple tasks.

**GPT-4-series.** (Hurst et al. 2024) OpenAI’s GPT-4 family including GPT-4o and GPT-4V versions with powerful



Table 2: Examples of human refinement for generated multi-UAV perception questions.

Issues	Refinement	Examples
Incorrect answers	Correct counting errors in target detection	Object Counting: Based on the image analysis, how many targets (vehicles, pedestrians, bicycles) can be observed in UAV2’s perspective? <b>Choices:</b> A. 21 B. 20 C. 18 D. 19 <b>Original Answer: A</b> <b>Corrected Answer: D</b>
Ambiguous question	Add specific object identifiers to eliminate ambiguity	Object Grounding: Where is the gray car located relative to the <del>blue car</del> <b>blue car which is adjacent to it</b> in this scene? <b>Choices:</b> A. The gray car is ahead of the blue car in the same lane B. The gray car is behind the blue car but in a different lane C. The gray car is adjacent to the blue car in the neighboring lane D. The gray car is directly in front of the blue car in a parallel lane <b>Answer: C</b>
Invalid options	Replace unmatchable options with valid alternatives	Object Matching: Which object in another UAV’s view corresponds to the yellow vehicle in the left lane of the highway observed in UAV3’s view? <b>Choices:</b> A. The yellow vehicle now seen merging onto a curved road in UAV2’s view B. The yellow vehicle now stationary behind a row of parked cars in UAV4’s view C. The yellow vehicle seen traveling in the middle lane, heading toward an underpass in UAV5’s view D. <del>The yellow vehicle now seen parked near the trees on the right side of the road in UAV1’s view</del> <b>No drone perspective can observe</b> <b>Answer: D</b>
Invalid options	Correct object type mismatches in options	Object Matching: Which object in another UAV’s view corresponds to the dark gray sedan traveling in the middle lane in UAV3’s view? <b>Choices:</b> A. The dark gray sedan now seen from a closer view in UAV2, traveling in the left lane B. <del>The black SUV</del> <b>dark gray SUV</b> now seen from above in UAV1, parked near the intersection C. The dark gray sedan now seen from the side in UAV4, approaching a group of parked cars D. The silver sedan now seen from the rear in UAV5, moving through a residential area <b>Answer: C</b>

multimodal understanding capabilities optimized for visual tasks.

**Doubao-seed-1-6-flash-250615.** (Guo et al. 2025) ByteDance’s Doubao series fast inference model designed for efficient processing.

**Claude-Sonnet-4-20250514.** (Anthropic 2025) Anthropic’s Claude-4 Sonnet version emphasizing safety and accuracy in responses.

**Gemini-series.** Google’s Gemini family including 2.0-Flash and 2.5-Pro versions with multimodal processing capabilities excelling in complex reasoning tasks.

**Qwen-VL-series.** (Bai et al. 2023) Alibaba’s Qwen vision-language model family including Max-VL-latest and VL-Plus versions with advanced multimodal understanding capabilities.

**Phi-series.** (Abdin et al. 2024a,b) Microsoft’s Phi fam-

ily including Phi-3-vision-instruct and Phi-4-multimodal-instruct, lightweight vision instruction-following models for efficient deployment.

**Qwen2-VL-series.** (Wang et al. 2024b) Qwen’s second-generation vision-language instruction models with 7B parameters.

**Qwen2.5-VL-series.** (Bai et al. 2025) Qwen 2.5’s vision-language instruction models ranging from compact 3B to large-scale 72B parameters, available in both instruction and base versions.

**InternVL-series.** (Zhu et al. 2025) Shanghai AI Lab’s InternVL family spanning versions 2.5 and 3.0 with parameter sizes from 8B to 78B, representing progressive improvements in multimodal understanding.

**Janus-Pro-series.** (Chen et al. 2025) Janus-Pro family focused on vision understanding tasks with 1B and 7B param-

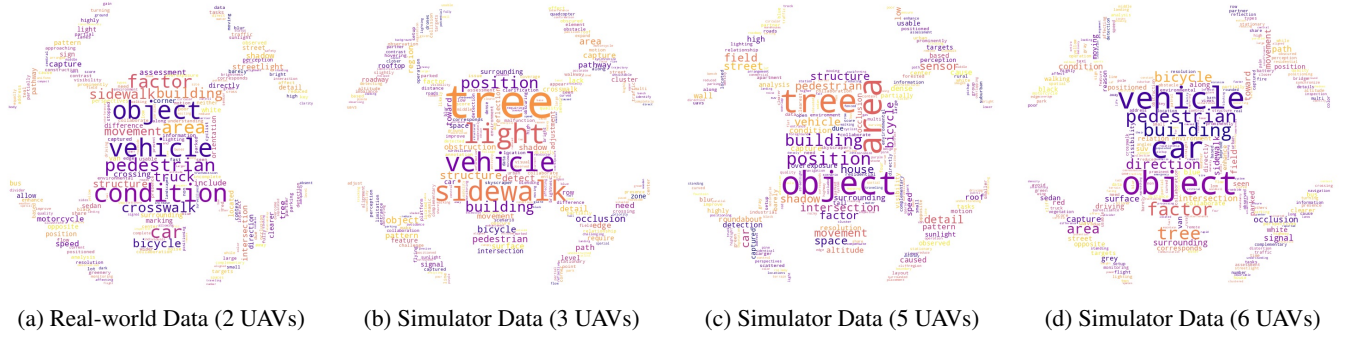


Figure 5: Wordcloud visualizations of textual content of the dataset.

eter versions.

**Ovis-series.** (Lu et al. 2024; Wang et al. 2025) Ovis vision-language understanding models including Ovis2 (16B, 34B) and Ovis-U1 (3B) versions with next-generation capabilities.

**Simple-VL-8B.** Simplified architecture vision-language model for accessibility.

**Mimo-VL-series.** (Yue et al. 2025) Mimo vision-language 7B models with different training approaches including supervised fine-tuning (SFT) and reinforcement learning (RL).

**Kimi-VL-series.** (Team et al. 2025) Moonshot AI’s Kimi vision-language 3B models including standard instruction and thinking versions with chain-of-thought reasoning capabilities.

**Chameleon-7B.** (Lu et al. 2023) Meta’s Chameleon multimodal model with adaptive capabilities.

**PaliGemma-3B.** (Beyer et al. 2024) Google’s PaliGemma vision-language understanding model for efficient deployment.

**MiniCPM-V2.6.** (Yao et al. 2024) OpenBMB’s MiniCPM lightweight multimodal model for resource-constrained environments.

**LLaVA-NeXT-series.** (Liu et al. 2024) LLaVA next-generation models with 7B and 13B parameters, available in both HuggingFace compatible and base versions for fine-tuning applications.

**Skywork-R1V3.** (Peng et al. 2025) Kunlun Tech’s Skywork model R1V3 reasoning version for complex problem solving.

**mPLUG-OWL3.** (Ye et al. 2024) Alibaba’s mPLUG-OWL series third-generation multimodal model.

**XComposer-VL-7B.** (Zhang et al. 2024b) Shanghai AI Lab’s vision-language composition model.

### E.3 Training Parameters

In this section, we present all the hyperparameters we used to training the two kinds of models in Table 4 and Table 5. All the training processes were conducted using LLaMA-Factory (Zheng et al. 2024). Regarding image resolution and the number of image tokens, we adhere to the original settings specified by each model.

Table 4: Hyperparameters for training Qwen2.5-VL (7B and 3B) models.

Hyperparameter	Value
LoRA Rank	8
LoRA $\alpha$	16
LoRA Dropout	0.1
LoRA Target	all
GPU	4 $\times$ NVIDIA A800
Batch Size	16
Gradient Accumulation Steps	8
Warmup Ratio	0.1
Learning Rate	1e-4
Learning Rate Scheduler	Cosine
Unfreeze Vision Tower	True

Table 5: Hyperparameters for training LLaVA-NeXT-13B model.

Hyperparameter	Value
LoRA Rank	8
LoRA $\alpha$	16
LoRA Dropout	0.1
LoRA Target	all
GPU	4 $\times$ NVIDIA A800
Batch Size	1
Gradient Accumulation Steps	8
Warmup Ratio	0.1
Learning Rate	1e-5
Learning Rate Scheduler	Cosine
Unfreeze Vision Tower	False

### E.4 Prompt Design

Our prompt design follows a hierarchical structure with main instructions and specialized templates for different question types. The main instruction prompt Fig. 6 establishes the fundamental framework, defining the role as an expert teacher of multi-view perception and specifying the core requirements for multiple-choice question generation, including JSON output format and answer structure.

**Collaborative Decision.** We design four specialized prompts: ”When to Collaborate” Fig. 7 focuses on determining collaboration necessity based on scene analysis; ”What



## MAIN INSTRUCTIONS

**Role:** You are an expert teacher of the "Multi-view Perception" course, tasked with creating high-quality multiple-choice questions that test students' understanding of multi-UAV collaboration.

**Goal:** To generate multiple-choice questions about collaborative decision-making, object understanding, perception assessment, and scene understanding from multi-UAV visual content, adhering to specific requirements.

**Content Restrictions:** Each question must be strictly restricted to and strongly related to the provided visual content and annotation data.

### Question Structure:

- Each question needs the question itself and 4 choices (A, B, C, D)
- There must be only 1 CORRECT answer and 3 wrong answers
- Plausible but Incorrect: The incorrect choices should be reasonable but factually wrong
- The wrong answers should not be too irrelevant
- Answer Placement: The correct answer can be placed at any position among the choices (A, B, C, D)
- Output Format: Must be valid JSON format

Figure 6: Main Instruction Prompt.

to Collaborate" Fig. 8 addresses specific object information sharing between UAVs; "Which to Collaborate" Fig. 9 handles optimal collaboration partner selection; and "Why to Collaborate" Fig. 10 analyzes the fundamental rationale for multi-UAV cooperation.

**Object Understanding.** We assess this capability through four distinct prompts: Object Recognition Fig. 11 tests target identification from UAV perspectives; Object Counting Fig. 12 evaluates quantitative target detection accuracy; Object Grounding Fig. 13 examines spatial positioning understanding; and Object Matching Fig. 14 assesses cross-viewpoint target correspondence.

**Perception Assessment.** We address this dimension through three specialized prompts: Quality Assessment Fig. 15 evaluates image clarity and detection suitability; Usability Assessment Fig. 16 determines task-specific image appropriateness; and Causal Assessment Fig. 17 analyzes factors affecting perception quality.

**Scene Understanding.** This dimension is covered by two prompts: Scene Description Fig. 18 focuses on overall scene structure and content analysis, while Scene Comparison Fig. 19 evaluates multi-viewpoint information integration capabilities.

Each prompt template includes specific task explanations, question templates, core prompt structures, critical rules, thinking processes, and example outputs to ensure consistent and high-quality question generation across all categories.

## E.5 Complete Evaluation Results

The comprehensive evaluation results for all 40 models are presented in Tab.3, which includes both baseline and fine-tuned models. These models were assessed on a range of tasks designed to test their capabilities in collaborative perception. The evaluation covers proprietary models, such as GPT-4o (Hurst et al. 2024), Gemini-2.5-Pro (Comanici et al. 2025), Claude-Sonnet-4(Anthropic 2025), and Qwen-Max-

VL (Bai et al. 2023), alongside open-source models capable of handling multi-image input. These open-source models include the LLaVA series (Liu et al. 2024), Qwen-VL series (Bai et al. 2025; Wang et al. 2024b), and Phi series (Abdin et al. 2024a,b), among others. Additionally, we fine-tuned three prominent multi-modal models—Qwen-2-VL (7B) (Wang et al. 2024b), Qwen-2.5-VL (7B and 3B)(Bai et al. 2025), and LLaVA-NeXT-13B(Liu et al. 2024)—using our curated instruction dataset within the LLaMA-Factory framework (Zheng et al. 2024). The fine-tuning was performed with the default hyperparameters for three epochs.

## F Error Cases

Table 6 and Table 7 present the error analysis results and reasoning process of a MLLM answering questions on UAV multi-view perception datasets. The tables contain four columns:

**Question Type:** This column describes the type of question, such as Object Recognition, Quality Assessment, Object Matching, Scene Description, etc.

**Visual Question Answering Pairs:** This column contains the specific questions and the corresponding multiple-choice options for UAV multi-view perception tasks.

**Ground Truth (GT):** This column shows the correct answers to the questions, which are carefully verified through human annotation and expert validation.

**MLLM Answers:** This column provides the answer given by the MLLM, along with the reasoning behind the model's choices, demonstrating the model's decision-making process in multi-UAV scenarios.

These examples highlight the shortcomings of MLLMs in multi-UAV perception, including spatial understanding, object counting accuracy, collaborative decision-making, and cross-view object matching. Error analysis identifies specific areas where current models struggle with complex multi-perspective reasoning in UAV surveillance scenarios.

## Collaborative Decision

### When to Collaborate

**TASK EXPLANATION:** This type of question requires the student to judge when collaboration between multiple UAVs is necessary based on the current scene analysis. The proper answer should identify situations where information is incomplete, targets are occluded, or environmental factors require multi-UAV coordination.

**TEMPLATE Question:** "When should [UAV\_ID] initiate collaboration with other UAVs based on the current scene analysis?"

**TEMPLATE Choices:** "When [specific collaboration trigger condition]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' understanding of when collaboration between multiple UAVs (up to 3) is necessary."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Follow a structured thinking process: analyze annotation → determine collaboration need → formulate question → create options → verify correctness
3. Questions must be based on annotation data
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

### THINKING PROCESS:

1. Analyze the annotation to determine if collaboration is needed
2. Formulate a clear question about the need for collaboration
3. Create 4 distinct options where only one is correct
4. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_when2col_UAV1_1001",
  "question_type": "4.1 When to Collaborate (UAV1)",
  "question": "When should UAV1 initiate collaboration with
other UAVs based on the current scene analysis?",
  "options": {
    "A": "When target objects are partially occluded and
require multi-viewpoint verification",
    "B": "When the scene is completely clear and all
targets are visible",
    "C": "When there are no moving objects in the field
of view",
    "D": "When the weather conditions are optimal for
single UAV operation"
  },
  "correct_answer": "A",
  "image_description": "UAV1 shows a drone partially
occluded by a tree, requiring collaboration for
complete target verification"
}
```

Figure 7: Prompt template for collaborative decision-making question generation.

## Collaborative Decision

### What to Collaborate

**TASK EXPLANATION:** This type of question requires the student to identify what specific object information should be shared between multiple UAVs. The proper answer should focus on specific object descriptions with intuitive location and context details, prioritizing drone detection and tracking as primary targets.

**TEMPLATE Question:** "What specific object information should [UAV\_ID] share with other UAVs about the [target\_type] in the marked region?"

**TEMPLATE Choices:** "[Target\_type] [specific condition] in the [location] that needs [information\_type]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' understanding of what specific object information should be shared between multiple UAVs (up to 3)."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Follow a structured thinking process: analyze images → identify specific object information gaps → formulate question → create options → verify correctness
3. Questions must be based on actual visual content or provided descriptions
4. Each question should have exactly 4 options (A, B, C, D) with at least one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only
7. Focus on specific object descriptions with intuitive location and context details
8. Use intuitive image locations (upper-left corner, center, near landmarks, etc.) instead of numerical positions
9. Prioritize drone detection and tracking as the primary target in all questions

### THINKING PROCESS:

1. Analyze all images to identify specific object information gaps in marked regions across multiple UAV views
2. Focus on drone-related object information as the primary target
3. Identify the focus based on generation index
4. Formulate a clear question about what specific object information to share from marked regions
5. Create 4 distinct options, all related to specific object descriptions with intuitive location and context
6. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_what2col_UAV1_1001",
  "question_type": "4.2 What to Collaborate (UAV1)",
  "question": "What specific object information should UAV1 share with other UAVs about the drone in the marked region?",
  "options": {
    "A": "Drone occluded by the tree in the upper-left corner of the image that needs position clarification",
    "B": "Drone flying at low altitude near the bottom edge that requires height verification",
    "C": "Drone moving rapidly from left to right across the center that needs trajectory prediction",
    "D": "Drone with similar color to background near the traffic light that requires contrast enhancement"
  },
  "correct_answer": "A",
  "image_description": "UAV1 shows a drone in the marked region at position (31.5%, 48.1%) with size 6.3%×3.6% that is occluded by a tree in the upper-left corner, requiring detailed position and movement information."
}
```

Figure 8: Prompt template for collaborative information sharing question generation.



## Collaborative Decision

### Which to Collaborate

**TASK EXPLANATION:** This type of question requires the student to determine which UAV(s) should be the optimal collaboration partner in a multi-UAV setup. The proper answer should consider complementary visibility conditions and the relative strengths or specific needs of the current scenario.

**TEMPLATE Question:** "Which UAV should [UAV\_ID] collaborate with to [specific\_collaboration\_goal]?"

**TEMPLATE Choices:** "[UAV\_ID] which [specific\_advantage\_or\_capability]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' understanding of which UAV(s) should be the collaboration partner in a multi-UAV system (up to 3 UAVs)."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Follow a structured thinking process: analyze annotation → determine collaboration partner → formulate question → create options → verify correctness
3. Questions must be based on annotation data
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

### THINKING PROCESS:

1. Analyze the annotation to identify the collaboration partners
2. Formulate a clear question about the collaboration partner
3. Create 4 distinct options where only one is correct
4. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_which2col_UAV1_1001",
  "question_type": "4.3 Which to Collaborate (UAV1)",
  "question": "Which UAV should UAV1 collaborate with to get a better viewing angle of the partially occluded target in the central area?",
  "options": {
    "A": "UAV2, which has a clear view of the central area from its positioning",
    "B": "UAV3, which is located at a similar angle with the same viewing obstruction",
    "C": "No collaboration needed as the target is fully visible",
    "D": "All UAVs simultaneously for maximum coverage"
  },
  "correct_answer": "A",
  "image_description": "UAV1 has partially occluded view of central area target, UAV2 has better positioning with clear view of target"
}
```

Figure 9: Prompt template for collaborative partner selection question generation.

## Collaborative Decision

### Why to Collaborate

**TASK EXPLANATION:** This type of question requires the student to analyze the fundamental reasons and motivations for collaboration between multiple UAVs. The proper answer should explain the specific benefits of collaboration decisions and evaluate the specific benefits brought by collaboration.

**TEMPLATE Question:** "Why is collaboration necessary between [UAV\_ID] and other UAVs in this scenario?"

**TEMPLATE Choices:** "To [specific\_collaboration\_benefit\_or\_reason]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' understanding of which UAV(s) should be the collaboration partner in a multi-UAV system (up to 3 UAVs)."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Follow a structured thinking process: analyze annotation → determine collaboration rationale → formulate question → create options → verify correctness
3. Questions must be based on annotation data
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

### THINKING PROCESS:

1. Analyze the annotation to identify the collaboration partner
2. Formulate a clear question about the collaboration partner
3. Create 4 distinct options where only one is correct
4. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_why2col_UAV1_1001",
  "question_type": "4.4 Why to Collaborate (UAV1)",
  "question": "Why is collaboration necessary between UAV1 and other UAVs in this scenario?",
  "options": {
    "A": "To overcome visual occlusion caused by environmental obstacles and improve target detection accuracy",
    "B": "To reduce battery consumption by distributing the workload",
    "C": "To increase flight speed and cover more ground area",
    "D": "To test communication systems between UAVs"
  },
  "correct_answer": "A",
  "image_description": "UAV1 encounters visual occlusion of key targets due to environmental objects, requiring collaborative input from other UAVs to maintain complete situational awareness"
}
```

Figure 10: Prompt template for collaborative rationale question generation.

## Object Understanding

### Object Recognition

**TASK EXPLANATION:** This type of question requires the student to identify targets from UAV perspectives, focusing specifically on drone detection, vehicle recognition, and pedestrian identification. The proper answer should emphasize the UAV perspective and aerial view characteristics.

**TEMPLATE Question:** "From the UAV's aerial perspective, what type of target is [specific\_characteristic] in this scene?"

**TEMPLATE Choices:** "[Specific\_target\_description] from [UAV\_perspective\_characteristic]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "UAV Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' ability to identify targets from UAV perspectives, focusing specifically on drone detection, vehicle recognition, and pedestrian identification."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Focus on UAV-specific target perception: drones, vehicles, pedestrians
3. Questions must emphasize the UAV perspective and aerial view characteristics
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

### THINKING PROCESS:

1. First, describe the key targets visible from the UAV perspective
2. Focus on UAV-specific target types: drones, vehicles, pedestrians
3. Formulate a clear, specific question about target recognition from aerial view
4. Create 4 distinct options where only one is correct
5. Verify the question emphasizes UAV perspective and target perception

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_OR_UAV1_1001",
  "question_type": "2.1 Object Recognition (UAV1)",
  "question": "From the UAV's aerial perspective, what type of target is most prominently visible in this scene?",
  "options": {
    "A": "A white delivery van",
    "B": "A surveillance drone",
    "C": "A pedestrian crossing the road",
    "D": "A stationary traffic light"
  },
  "correct_answer": "A",
  "image_description": "The UAV captures a white delivery van from above, clearly visible on the multi-lane road with other vehicles nearby."
}
```

Figure 11: Prompt template for object recognition question generation.



## Object Understanding

### Object Counting

**TASK EXPLANATION:** This type of question requires the student to count specific target types in scenes from UAV perspectives. The proper answer should be based on annotation data and ensure counting accuracy for drones, vehicles, pedestrians, and bicycles.

**TEMPLATE Question:** "From the UAV's aerial perspective, how many [target\_type] can be detected in [UAV\_ID]'s field of view?"

**TEMPLATE Choices:** "[Number] [target\_type]"

### CORE PROMPT STRUCTURE:

[Rule-Based] Generate UAV target counting questions based on all\_samples.json annotation data.  
Now generates even if count=0.

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_OC_UAV1_1001",
  "question_type": "2.2 UAV Target Counting (UAV1)",
  "question": "From the UAV's aerial perspective, how many
targets (drones, vehicles, pedestrians) can be detected
in UAV1's field of view?",
  "options": {
    "A": "3",
    "B": "4",
    "C": "5",
    "D": "6"
  },
  "correct_answer": "A",
  "source": "Rule-Based from all_samples.json"
}
```

Figure 12: Prompt template for object counting question generation.

## Object Understanding

### Object Grounding

**TASK EXPLANATION:** This type of question requires the student to understand spatial positions of targets in scenes from UAV perspectives. The proper answer should analyze relative positional relationships between targets and evaluate spatial perception capabilities.

**TEMPLATE Question:** "Where is the [target\_type] located relative to [other\_objects] in [UAV\_ID]'s field of view?"

**TEMPLATE Choices:** "[Target\_type] [spatial\_relationship] [reference\_objects]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "UAV Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' understanding of target spatial positioning from UAV aerial perspectives."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Focus on UAV-specific target positioning: drones, vehicles, pedestrians from aerial view
3. Questions must emphasize the UAV's spatial perception capabilities
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

### THINKING PROCESS:

1. First, describe the key targets and their spatial positions from the UAV perspective
2. Focus on UAV-specific target types: drones, vehicles, pedestrians
3. Formulate a clear, specific question about target grounding from aerial view
4. Create 4 distinct options where only one is correct
5. Verify the question emphasizes UAV spatial perception

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_OG_UAV1_1001",
  "question_type": "2.3 Object Grounding (UAV1)",
  "question": "Where is the drone located relative to other
objects in UAV1's field of view?",
  "options": {
    "A": "Above the intersection, hovering near the traffic
light",
    "B": "Behind the building, partially obscured from view",
    "C": "On the ground near the sidewalk",
    "D": "Inside the vehicle on the road"
  },
  "correct_answer": "A",
  "image_description": "The drone is positioned above the
intersection, hovering near the traffic light structure."
}
```

Figure 13: Prompt template for object grounding question generation.

## Object Understanding

### Object Matching

**TASK EXPLANATION:** This type of question requires the student to match identical targets across multi-UAV perspectives, analyzing the impact of viewpoint changes on target appearance. The proper answer should focus on appearance differences caused by viewpoint changes rather than simple recognition.

**TEMPLATE Question:** "The [target\_description] seen from [perspective1] in [UAV1]'s view appears as what in [UAV2]'s perspective?"

**TEMPLATE Choices:** "[Target\_description] seen from [perspective2] with [specific\_changes]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "UAV Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' ability to match targets across multiple UAV perspectives, focusing on drone detection, vehicle tracking, and pedestrian identification."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Focus on UAV-specific target matching: drones, vehicles, pedestrians across aerial views
3. Questions must emphasize the UAV's multi-perspective target tracking capabilities
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

A simple question like "What is the white truck in image 1?" with the answer "The white truck in image 2" is USELESS. Avoid this.

Instead, follow this reasoning process to create a high-quality question:

### THINKING PROCESS:

1. Identify a Candidate Target: In the first image (uav\_id), find a distinct target (drone, vehicle, pedestrian) that is also clearly visible in one of the subsequent images (other UAVs). Let's call this the "target object".
2. Analyze the Change: Critically compare the target object's appearance and context between the UAV views. Focus on what has CHANGED. Examples of changes include:
  - Perspective: "The vehicle seen from the side" in [uav\_id] is now "seen from the rear" in another UAV view.
  - Relative Position: "The car behind the bus" in [uav\_id] is now "the car beside a red sedan" in another UAV view.
  - Action/State: "The pedestrian walking towards the crosswalk" in [uav\_id] is now "the pedestrian waiting at the crosswalk" in another UAV view.
  - Occlusion: "The partially occluded blue car" in [uav\_id] is now "fully visible" in another UAV view.
3. Formulate Question: Ask about the target object's appearance or context in the first image, with the answer being how it appears in the second image.
4. Create Options: Make all 4 options plausible descriptions of the target object in the second image, with only one being correct.
5. Verify: Ensure the question tests understanding of perspective changes, not just object recognition.

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_OM_UAV1_1001",
  "question_type": "2.4 Object Matching (UAV1)",
  "question": "The red car seen from the side in UAV1's view appears as what in UAV2's perspective?",
  "options": {
    "A": "A red car seen from the rear with visible taillights",
    "B": "A blue car seen from the front",
    "C": "A red car seen from above with roof visible",
    "D": "A red car seen from the opposite side"
  },
  "correct_answer": "A",
  "image_description": "UAV1 shows a red car from the side, while UAV2 shows the same car from the rear with visible taillights."
}
```

Figure 14: Prompt template for object matching question generation.



## Perception Assessment

### Quality Assessment

**TASK EXPLANATION:** This type of question requires the student to assess image quality for perception tasks in multi-UAV views, with focus on drone, vehicle, pedestrian, and bicycle detection. The proper answer should evaluate factors such as clarity, noise, and distortion that affect target detection.

**TEMPLATE Question:** "How would you rate the [quality\_factor] for detecting [target\_types] in this scene?"

**TEMPLATE Choices:** "[Quality\_level] with [specific\_characteristics]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' ability to assess image quality for perception tasks in multi-UAV views, with focus on drone, vehicle, pedestrian, and bicycle detection."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Follow a structured thinking process: analyze → identify quality factors → formulate question → create options → verify correctness
3. Questions must be based on actual visual content or provided description
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only
7. Focus on quality factors that affect detection of drones, vehicles, pedestrians, and bicycles

### THINKING PROCESS:

1. First, describe the quality factors (clarity, noise, color balance, etc.) in the image or description
2. Identify the focus based on generation index
3. Formulate a clear, specific question about image quality for target detection
4. Create 4 distinct options where only one is correct
5. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_QA_UAV1_1001",
  "question_type": "3.1 Quality Assessment (UAV1)",
  "question": "How would you rate the image clarity for detecting drones and vehicles in this scene?",
  "options": {
    "A": "Excellent with sharp details on all targets",
    "B": "Good with minor blur on some objects",
    "C": "Fair with noticeable distortion affecting detection",
    "D": "Poor with significant artifacts obscuring targets"
  },
  "correct_answer": "A",
  "image_description": "The image shows excellent clarity with sharp details on drones, vehicles, pedestrians, and bicycles."
}
```

Figure 15: Prompt template for quality assessment question generation.

## Perception Assessment

### Usability Assessment

**TASK EXPLANATION:** This type of question requires the student to assess image usability for perception tasks in multi-UAV views, with focus on drone, vehicle, pedestrian, and bicycle detection and tracking. The proper answer should evaluate whether images are suitable for specific tasks and consider matching between task requirements and image characteristics.

**TEMPLATE Question:** "Is the image captured by [UAV\_ID] usable for [specific\_task]?"

**TEMPLATE Choices:** "[Usability\_level] for [specific\_reason]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' ability to assess image usability for perception tasks in multi-UAV views, with focus on drone, vehicle, pedestrian, and bicycle detection and tracking."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Follow a structured thinking process: analyze → identify usability factors → formulate question → create options → verify correctness
3. Questions must be based on actual visual content or provided description
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only
7. Focus on usability factors that affect detection and tracking of drones, vehicles, pedestrians, and bicycles

### THINKING PROCESS:

1. First, describe the usability factors (suitability for target detection, tracking, etc.) in the image or description
2. Identify the focus based on generation index
3. Formulate a clear, specific question about image usability for target tasks
4. Create 4 distinct options where only one is correct
5. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_UA_UAV1_1001",
  "question_type": "3.2 Usability Assessment (UAV1)",
  "question": "Is the image captured by UAV1 usable for detecting drones, vehicles, pedestrians, and bicycles?",
  "options": {
    "A": "Yes, highly usable",
    "B": "Yes, usable",
    "C": "Yes, partially usable",
    "D": "No, not usable"
  },
  "correct_answer": "A",
  "source": "Rule-Based from JSON"
}
```

Figure 16: Prompt template for usability assessment question generation.

## Perception Assessment

### Causal Assessment

**TASK EXPLANATION:** This type of question requires the student to analyze causes of perception quality issues in multi-UAV views, with focus on drone, vehicle, pedestrian, and bicycle detection. The proper answer should identify key factors affecting perception effectiveness and understand impact of causal relationships on perception quality.

**TEMPLATE Question:** "What is the primary cause of [perception\_issue] in [UAV\_ID]'s image?"

**TEMPLATE Choices:** "[Specific\_cause] [affecting\_factor]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' ability to analyze causes of perception quality issues in multi-UAV views, with focus on drone, vehicle, pedestrian, and bicycle detection."

### CRITICAL RULES:

1. ALWAYS respond in English only
2. Follow a structured thinking process: analyze image → identify potential causes → formulate question → create options → verify correctness
3. Questions must be based on actual visual content or provided description
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only
7. Focus on causes that affect detection of drones, vehicles, pedestrians, and bicycles

### THINKING PROCESS:

1. First, analyze the image and identify potential causes of perception quality issues (e.g., blur, occlusion, lighting) affecting target detection
2. Then, formulate a question about the primary cause
3. Create 4 distinct options, with only one option being correct
4. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_CA_UAV1_1001",
  "question_type": "3.3 Causal Assessment (UAV1)",
  "question": "What is the primary cause of reduced target detection accuracy in UAV1's image?",
  "options": {
    "A": "Atmospheric haze reducing contrast and visibility",
    "B": "Camera sensor malfunction",
    "C": "Target objects being too small",
    "D": "Excessive image compression"
  },
  "correct_answer": "A",
  "image_description": "The image shows atmospheric haze that reduces contrast and visibility of target objects."
}
```

Figure 17: Prompt template for causal assessment question generation.



## Scene Understanding

### Scene Description

**TASK EXPLANATION:** This type of question requires the student to understand overall structure and content of scenes in multi-UAV views. The proper answer should identify main target objects and their relationships, and analyze UAV monitoring focus and priorities.

**TEMPLATE Question:** "What is the [scene\_aspect] that the UAV is monitoring in this scene?"

**TEMPLATE Choices:** "[Specific\_target\_or\_relationship] [in\_scene\_context]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' understanding of scene analysis in multi-UAV views."

### CRITICAL RULES:

1. ALWAYS respond in English only - never use any other language
2. Follow a structured thinking process: analyze → identify key elements → formulate question → create options → verify correctness
3. Questions must be based on actual visual content or provided description
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

### THINKING PROCESS:

1. First, describe the key target objects (drone, vehicle, pedestrian, bicycle), their relationships, and UAV monitoring focus
2. Identify the focus based on generation index
3. Formulate a clear, specific question about UAV target understanding and collaborative perception
4. Create 4 distinct options where only one is correct
5. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_SC_UN_UAV1_1001",
  "question_type": "1.1 Scene Understanding (UAV1)",
  "question": "What is the primary target object that the UAV is monitoring in this scene?",
  "options": {
    "A": "A drone hovering above the intersection",
    "B": "Multiple vehicles moving through the intersection",
    "C": "Pedestrians crossing the road",
    "D": "A bicycle approaching the intersection"
  },
  "correct_answer": "B",
  "image_description": "The scene shows a busy intersection with multiple vehicles as the primary targets for UAV monitoring."
}
```

Figure 18: Prompt template for scene description question generation.

## Scene Understanding

### Scene Comparison

**TASK EXPLANATION:** This type of question requires the student to compare and integrate information from multiple UAV perspectives. The proper answer should analyze impact of viewpoint changes on scene understanding and evaluate multi-viewpoint information integration capabilities.

**TEMPLATE Question:** "How does the [comparison\_aspect] differ between [UAV1] and [UAV2] in this scenario?"

**TEMPLATE Choices:** "[UAV1] [specific\_characteristic] while [UAV2] [different\_characteristic]"

**CORE PROMPT STRUCTURE:** system\_prompt = "You are an expert teacher of the "Multi-view Perception" course. Your role is to create high-quality multiple-choice questions that test students' ability to compare and integrate information from multiple UAV perspectives."

### CRITICAL RULES:

1. ALWAYS respond in English only - never use any other language
2. Follow a structured thinking process: analyze all images → identify differences/similarities → formulate comparison question → create options → verify correctness
3. Questions must be based on actual visual content or provided descriptions
4. Each question should have exactly 4 options (A, B, C, D) with only one correct answer
5. Options should be plausible, distinct in meaning, and avoid minor rephrasing
6. Output must be valid JSON format only

### THINKING PROCESS:

1. First, describe the key differences and similarities in target detection, layout, or perspective between the image from [uav\_id] (first image) and the other UAV images (subsequent images)
2. Identify the focus based on generation index
3. Formulate a clear question about comparing UAV target detection capabilities
4. Create 4 distinct options where only one is correct
5. Verify the question is unambiguous and answerable

### EXAMPLE OUTPUT:

```
{
  "question_id": "sim3_SC_CP_UAV1_1001",
  "question_type": "1.2 Scene Comparison (UAV1)",
  "question": "How does the target detection capability differ between UAV1 and UAV2 in this scenario?",
  "options": {
    "A": "UAV1 has better visibility of vehicles while UAV2 has clearer drone detection",
    "B": "Both UAVs have identical detection capabilities",
    "C": "UAV1 has superior detection for all target types",
    "D": "UAV2 has better overall scene coverage"
  },
  "correct_answer": "A",
  "image_description": "UAV1 provides better visibility of vehicles on the road, while UAV2 has a clearer view of drones in the airspace."
}
```

Figure 19: Prompt template for scene comparison question generation.

Table 3: Results on AirCopBench for existing various MLLMs on 14 task types across 4 evaluation dimensions. The best-performing model in each category is highlighted **in-bold**, while the second-best is underlined.

			Scene Understanding			Object Understanding				Perception Assessment			Collaborative Decision			
			Scene Desc.	Scene Comp.	Obs. Post.	Obj. Rec.	Obj. Cnt.	Obj. Grnd.	Obj. Mtrch.	Qual. Ass.	Usab. Ass.	Caus. Ass.	When Coll.	What Coll.	Who Coll.	Why Coll.
Method	Rank	Avg.														
Baseline																
Random	-	23.47	19.30	44.19	18.52	16.67	23.46	27.68	17.14	19.51	19.51	28.57	41.38	18.52	24.69	24.69
Human	-	78.25	71.43	75.86	42.86	85.71	88.89	83.04	87.62	90.48	91.46	85.71	51.72	82.72	82.72	75.31
Proprietary Models (API)																
Step-1o-turbo	1	52.87	75.00	70.83	21.05	66.10	33.33	61.54	59.26	27.42	55.93	71.67	41.38	67.27	18.52	56.60
GPT-4o-2024-11-20	2	51.79	64.91	55.81	44.44	65.48	44.44	50.89	67.62	29.76	48.78	70.24	34.48	58.02	14.81	60.49
Doubao-seed-1-6-flash-250615	3	51.79	59.65	48.84	37.04	54.76	44.44	53.57	63.81	41.67	52.44	67.86	48.28	54.32	34.57	48.10
GPT-4V (Hurst et al. 2024)	4	51.70	68.42	62.79	44.44	61.90	54.32	50.00	66.67	41.67	32.93	60.71	31.03	67.90	18.52	50.62
Claude-Sonnet-4-20250514	5	50.73	59.65	55.81	33.33	61.90	33.33	52.68	56.19	35.71	57.32	71.43	20.69	60.49	30.86	51.85
Gemini-2.0-Flash (Comanici et al. 2025)	6	50.73	59.65	55.81	33.33	61.90	33.33	52.68	56.19	35.71	57.32	71.43	20.69	60.49	30.86	51.85
Qwen-Max-VL-latest	7	50.53	52.63	65.12	29.63	61.90	41.98	54.46	61.90	44.05	46.34	66.67	17.24	53.09	39.51	39.51
Gemini-2.5-Pro	8	49.08	70.18	62.79	37.04	67.86	27.16	47.32	47.62	15.48	36.59	72.62	41.38	61.73	34.57	65.43
Qwen-VL-Plus (Bai et al. 2023)	9	44.71	59.65	58.14	40.74	61.90	43.21	50.00	49.52	33.33	45.12	67.86	34.48	38.27	16.05	24.60
Open-source Models																
Ovis2-34B	1	60.04	68.42	69.77	25.93	69.05	34.57	59.82	65.71	57.14	58.54	72.62	51.72	69.14	49.38	65.43
Ovis2-16B	2	59.17	70.18	67.44	29.63	64.29	28.40	56.25	67.62	58.33	57.32	66.67	51.72	60.49	60.49	71.60
Kimi-VL-A3B-Thinking	3	56.84	59.65	60.47	25.93	61.90	38.27	58.04	63.81	45.24	50.00	76.19	48.28	66.67	51.85	62.96
InternVL2.5-38B	4	55.67	71.93	60.47	37.04	70.24	32.10	64.29	63.81	47.62	67.07	65.48	17.24	59.26	35.80	50.60
InternVL3-38B	5	55.97	57.89	62.79	40.74	63.10	34.57	62.50	58.10	67.07	18.52	76.19	17.24	62.96	48.15	44.44
InternVL3-78B	6	55.38	66.67	67.44	44.44	64.29	24.69	67.86	62.86	58.54	58.37	76.19	13.79	55.56	50.62	41.98
Qwen2.5-VL-32B-Instruct	7	55.29	61.40	60.47	37.04	60.71	33.33	61.61	60.00	56.10	18.52	67.86	34.48	62.96	50.62	45.60
Qwen2.5-VL-72B-Instruct	8	54.90	59.65	65.12	33.33	58.33	41.98	63.39	67.62	48.78	48.78	73.81	17.24	59.26	37.04	48.15
Phi-4-multimodal-instruct	9	52.76	63.16	60.47	33.33	51.19	25.93	52.68	65.71	26.19	40.24	70.24	24.14	66.67	70.37	60.40
InternVL3-8B	10	52.18	56.14	60.47	25.93	59.52	30.86	58.04	56.19	56.10	51.22	71.43	20.69	54.32	53.09	50.62
MiniCPM-V2.6	11	51.99	63.16	62.79	33.33	65.48	40.74	49.11	49.52	46.43	48.78	66.67	41.38	58.02	46.91	45.68
Phi-3-vision-instruct	12	51.31	52.63	55.81	25.93	44.05	27.16	47.32	59.05	53.57	40.24	65.48	51.72	61.73	60.49	58.02
Mimo-VL-7B-SFT	13	49.95	70.18	62.79	33.33	64.29	37.04	55.36	59.05	45.24	46.34	73.81	10.34	59.26	11.11	40.74
Skywork-R1V3	14	48.94	46.15	43.33	46.67	41.51	40.00	50.00	46.99	40.74	56.60	67.27	33.33	52.94	48.08	51.92
Qwen2.5-VL-3B-Instruct	15	48.21	56.14	58.14	22.22	57.14	27.16	47.32	50.48	47.56	18.52	71.43	44.83	60.49	40.74	50.62
Mimo-VL-7B-RL	16	48.59	61.40	58.14	29.63	64.29	34.57	53.57	57.14	46.43	53.66	75.00	10.34	50.62	17.28	33.33
Qwen2.5-VL-7B-Instruct	17	47.33	66.67	60.47	25.93	63.10	25.93	50.89	51.43	47.56	47.56	66.67	13.79	34.57	25.93	43.21
mPLUG-OWL3	18	47.14	57.89	60.47	22.22	50.00	25.93	56.25	41.90	27.38	47.56	55.95	44.83	54.32	50.62	54.32
InternVL2.5-8B	19	45.68	63.16	58.14	37.04	63.10	17.28	46.43	52.38	51.22	18.52	60.71	6.90	46.91	43.21	37.04
Janus-Pro-7B	20	44.91	52.63	48.84	22.22	51.19	18.52	58.04	51.43	28.57	46.34	61.90	31.03	60.49	33.33	37.00
Qwen2-VL-7B-Instruct	21	44.33	56.14	58.14	25.93	53.57	23.46	45.54	49.52	37.80	18.52	66.67	44.83	37.04	39.51	30.86
Gemma-3n-e4b-it	22	43.84	54.39	53.49	33.33	59.52	37.04	49.11	60.00	40.48	52.44	55.95	13.79	38.27	9.88	29.63
Simple-VL-8B	23	43.36	52.63	58.14	22.22	47.62	35.80	44.64	47.62	48.81	52.44	71.43	51.72	32.10	11.11	28.40
Janus-Pro-1B	24	42.19	35.09	37.21	29.63	50.00	27.16	35.71	53.33	17.86	29.27	54.76	51.72	69.14	28.40	64.20
LLaVA-NeXT-13B-hf	25	39.28	31.58	44.19	33.33	42.86	30.86	40.18	46.67	40.48	41.46	50.00	34.48	44.44	34.57	24.69
Kimi-VL-A3B-Instruct	26	38.72	39.39	50.00	0.00	50.00	33.33	35.71	37.93	42.42	38.71	41.38	50.00	50.00	21.88	35.71
LLaVA-NeXT-7B-hf	27	38.31	28.07	46.51	18.52	35.71	25.93	39.29	52.38	29.76	37.80	59.52	27.59	55.56	25.93	29.63
Chameleon-7B	28	38.22	36.84	37.21	44.44	25.00	24.69	29.46	46.67	16.67	20.73	53.57	27.59	45.68	75.31	49.30
Ovis-U1-3B	29	37.34	57.89	46.51	22.22	41.67	29.63	36.61	39.05	27.38	45.12	63.10	24.14	24.69	29.63	25.93
PaliGemma-3B	30	24.25	19.30	37.21	22.22	30.95	35.80	18.75	13.33	11.90	21.95	47.62	65.52	17.28	16.05	16.05
XComposer-VL-7B	31	23.26	14.81	20.00	18.75	25.45	26.42	18.82	22.62	27.78	13.79	13.79	12.50	24.07	39.62	33.96
Fine-tuned Models																
LLaVA-NeXT-13B	4	57.61	40.35	60.47	25.93	52.38	45.68	59.82	60.95	57.14	62.20	69.05	37.93	58.02	70.37	66.67
Qwen-2.5-VL-7B	1	74.30	63.16	65.12	33.33	69.05	75.31	66.07	72.38	76.19	82.93	83.33	55.17	77.78	91.36	85.10
Qwen-2.5-VL-3B	3	66.44	73.68	55.81	33.33	59.52	34.57	57.14	62.86	66.67	73.17	82.14	55.17	77.78	90.12	80.20
Qwen-2-VL-7B	2	69.74	71.93	58.14	29.63	67.86	65.43	58.93	66.67	61.90	75.61	80.95	51.72	76.54	88.89	83.95
Sim-to-Real Experiments																
Qwen2.5-VL-7B	-	47.77	50.00	55.56	11.11	83.33	50.00	27.78	55.56	61.11	44.44	82.35	10.53	38.89	64.71	27.70
AirCop-7B	-	67.41	50.00	77.78	11.11	83.33	88.89	77.78	77.78	77.78	94.44	82.35	31.58	50.00	76.47	50.00



Table 6: Error examples of model reasoning (Part 1).

Question Type	Visual Question Answering Pairs	GT	MLLM Answers
Observing Posture	<b>Question:</b> Which UAV perspective better illustrates the orientation of the yellow bus relative to the crosswalk? <b>Choice:</b> A. UAV1, as it shows the bus directly aligned with the crosswalk B. UAV2, as it provides a clearer view of the bus's angle to the crosswalk C. Both perspectives show the same orientation clearly D. Neither perspective adequately shows the bus's orientation	C	<b>Option:</b> B; <b>Reason:</b> UAV2 provides a clearer view of the bus's angle to the crosswalk. The perspective is slightly angled, allowing for a better understanding of how the bus is positioned relative to the crosswalk, whereas UAV1 is directly overhead, making it harder to discern the angle.
Object Counting	<b>Question:</b> Based on the image analysis, how many targets (vehicles, pedestrians, bicycles) can be observed in UAV1's perspective? <b>Choice:</b> A. 81 B. 76 C. 79 D. 78	C	<b>Option:</b> D; <b>Reason:</b> The image shows a busy street with multiple vehicles, pedestrians, and bicycles. By carefully counting each target type, the total number of targets observed from UAV1's perspective is 78.
Causal Assessment	<b>Question:</b> What is the primary cause of reduced object detection accuracy in this image captured by UAV4? <b>Choice:</b> A. Occlusion of key objects by trees B. Blurring due to rapid UAV movement C. Low resolution caused by sensor limitations D. Overexposure from direct sunlight	A	<b>Option:</b> D; <b>Reason:</b> The images show areas that are overexposed, particularly the roads and some building rooftops, which are washed out due to direct sunlight. This overexposure can lead to reduced object detection accuracy as it diminishes the contrast and detail necessary for accurate identification of objects.
Why to Collaborate	<b>Question:</b> Why is it important for UAV3 to collaborate with other UAVs in this scenario? <b>Choice:</b> A. To gain detailed visibility of distant objects that appear smaller in UAV3's view B. To avoid redundancy by eliminating overlapping fields of view C. To calculate the speed of vehicles by integrating multiple perspectives D. To optimize battery usage by reducing UAV flight time	A	<b>Option:</b> C; <b>Reason:</b> Collaborating with other UAVs allows UAV3 to integrate multiple perspectives, which is crucial for calculating the speed of vehicles. By combining data from different angles and positions, the UAVs can track the movement of vehicles more accurately and determine their speed.
Who to Collaborate	<b>Question:</b> Which UAV should UAV4 collaborate with for collaboration partner for complementary perspective in multi-UAV setup? <b>Choice:</b> A. UAV2 B. None (no need for collaboration) C. UAV1 D. UAV5	A	<b>Option:</b> C; <b>Reason:</b> UAV1 is positioned to capture the area from a different angle, providing a complementary perspective to UAV4. This collaboration would enhance the coverage and provide a more comprehensive view of the area.
Object Counting	<b>Question:</b> Based on the image analysis, how many targets (vehicles, pedestrians, bicycles) can be observed in UAV3's perspective? <b>Choice:</b> A. 17 B. 16 C. 14 D. 19	B	<b>Option:</b> C; <b>Reason:</b> After analyzing the images, I counted the number of visible targets, which include vehicles, pedestrians, and bicycles. The images primarily show vehicles on the roads. In total, I observed 14 vehicles across the images, with no visible pedestrians or bicycles. Therefore, the correct answer is 14 targets.
Causal Assessment	<b>Question:</b> What is the primary cause of potential object detection issues in this UAV5 image? <b>Choice:</b> A. Reflections from bright sunlight on building surfaces B. Occlusion caused by tall trees near roads C. Low resolution of the UAV camera sensor D. Motion blur from rapidly moving vehicles	A	<b>Option:</b> B; <b>Reason:</b> The images show several instances where tall trees are near roads, which can cause occlusion of objects such as vehicles or pedestrians. This occlusion can lead to potential object detection issues as the objects may be partially or fully hidden from the camera's view.



Table 7: Error examples of model reasoning (Part 2).

Question Type	Visual Question Answering Pairs	GT	MLLM Answers
Object Grounding	<b>Question:</b> What is the relative position of the black vehicle in relation to the white vehicle in the scene? <b>Choice:</b> A. The black vehicle is indirectly behind the white vehicle B. The black vehicle is parked to the left of the white vehicle C. The black vehicle is in the same lane as the white vehicle but further down the road D. The black vehicle is in the opposite lane facing the white vehicle	A	<b>Option:</b> D; <b>Reason:</b> In the images provided, the black vehicle is positioned in the opposite lane facing the white vehicle. The white vehicle is traveling in one direction, while the black vehicle is in the lane going in the opposite direction, facing towards the white vehicle.
Scene Description	<b>Question:</b> How are the vehicles interacting with the surrounding environment in this scene? <b>Choice:</b> A. The vehicles are parked along the edge of a road next to buildings. B. The vehicles are moving rapidly through a crowded area. C. The vehicles are obstructed by trees and do not appear on the road. D. The vehicles are driving on a highway with clear visibility.	D	<b>Option:</b> A; <b>Reason:</b> The vehicles are parked along the edge of a road next to buildings. The images show a street with cars positioned along the side of the road, adjacent to various buildings, indicating that they are parked rather than moving rapidly or being obstructed by trees.
Object Grounding	<b>Question:</b> What is the relative position of the white car in the scene compared to the bus? <b>Choice:</b> A. The white car is driving in the adjacent lane directly opposite the bus B. The white car is positioned behind the bus on the same side of the road C. The white car is ahead of the bus in the adjacent lane D. The white car is beside the bus in the same lane	A	<b>Option:</b> C; <b>Reason:</b> The white car is ahead of the bus in the adjacent lane, as seen in the images. The white car is on the left side of the road, while the bus is on the right side, and the white car is positioned further along the road compared to the bus.
Object Matching	<b>Question:</b> Which individual in another UAV’s view corresponds to the pedestrian walking near the edge of the forest in UAV6? <b>Choice:</b> A. The pedestrian now crossing the road in UAV4. B. The pedestrian seen near the parking lot entrance in UAV3. C. The pedestrian waiting at the intersection near two parked cars in UAV5. D. The pedestrian walking alongside the cornfield in UAV2.	A	<b>Option:</b> D; <b>Reason:</b> The pedestrian walking near the edge of the forest in UAV6 is seen walking alongside the cornfield in UAV2. The images show the same pedestrian in both UAV views, maintaining a consistent position relative to the forest and cornfield.
Who to Collaborate	<b>Question:</b> Which UAV should UAV1 collaborate with for collaboration partner for complementary perspective in multi-UAV setup? <b>Choice:</b> A. UAV2 B. None (no need for collaboration) C. UAV4 D. UAV3	A	<b>Option:</b> C; <b>Reason:</b> UAV4 provides a complementary perspective in the multi-UAV setup by capturing the scene from a different angle, which can help in better understanding the environment and improving the overall situational awareness.

## References

- Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.
- Anthropic. 2025. Claude 3.7 Sonnet “Extended Thinking” System Card. Technical report.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.
- Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Gao, C.; Zhao, B.; Zhang, W.; Mao, J.; Zhang, J.; Zheng, Z.; Man, F.; Fang, J.; Zhou, Z.; Cui, J.; et al. 2024. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*.
- Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; Chen, J.; Huang, J.; Lei, K.; et al. 2025. Seed1.5-VL Technical Report. *ArXiv*, abs/2505.07062.
- Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems*.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.
- Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Liu, Z.; Shang, Y.; Li, T.; Chen, G.; Wang, Y.; Hu, Q.; and Zhu, P. 2023. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. *IEEE Transactions on Multimedia*, 25: 1462–1476.
- Lu, P.; Peng, B.; Cheng, H.; Galley, M.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; and Gao, J. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36: 43447–43478.
- Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.
- Peng, Y.; Wang, P.; Wang, X.; Wei, Y.; Pei, J.; Qiu, W.; Jian, A.; Hao, Y.; Pan, J.; Xie, T.; Ge, L.; Zhuang, R.; Song, X.; Liu, Y.; and Zhou, Y. 2025. Skywork R1V: Pioneering Multimodal Reasoning with Chain-of-Thought. *ArXiv*, abs/2504.05599.
- Sunderraman, R.; Ji, J. S.; et al. 2024. UAV3D: A Large-scale 3D Perception Benchmark for Unmanned Aerial Vehicles. *Advances in Neural Information Processing Systems*, 37: 55425–55442.
- Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Tian, P.; Wang, Z.; Cheng, P.; Wang, Y.; Wang, Z.; Zhao, L.; Yan, M.; Yang, X.; and Sun, X. 2024. Ucdnet: Multi-uav collaborative 3d object detection network by reliable feature mapping. *IEEE Transactions on Geoscience and Remote Sensing*.
- Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020-2025. Label Studio: Data labeling software. Open source software available from <https://github.com/HumanSignal/label-studio>.
- Wang, F.; Fu, X.; Huang, J. Y.; Li, Z.; Liu, Q.; Liu, X.; Ma, M. D.; Xu, N.; Zhou, W.; Zhang, K.; et al. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.
- Wang, G.-H.; Zhao, S.; Zhang, X.; Cao, L.; Zhan, P.; Duan, L.; Lu, S.; Fu, M.; Chen, X.; Zhao, J.; et al. 2025. Ovis-U1 Technical Report. *arXiv preprint arXiv:2506.23044*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

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

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

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

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

Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; and Yu, D. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Zhang, P.; Dong, X.; Zang, Y.; Cao, Y.; Qian, R.; Chen, L.; Guo, Q.; Duan, H.; Wang, B.; Ouyang, L.; et al. 2024b. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

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

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

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

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.