

# OneScene-Bench: Do Vision-Language Models Integrate Evidence Across Views?

Swathi Suhas Anurag Das Bernt Schiele Jonas Fischer

Max Planck Institute for Informatics,  
Saarland Informatics Campus, Saarbrücken, Germany

## Abstract

Vision-language models (VLMs) increasingly process multiple views of the same scene, but it remains unclear whether they integrate these observations into a coherent scene-level representation. Existing multi-image evaluations do not isolate this capability, as tasks may admit correct solutions through independent per-view processing without requiring cross-view correspondence. We introduce **OneScene-Bench**, a controlled benchmark designed to explicitly distinguish when cross-view integration is necessary. Our benchmark evaluates three complementary capabilities on real-world multi-view scenes: unique object counting, cross-view consistency, and spatial reasoning. For counting, we control object overlap across views to separate cases solvable by simple per-view aggregation from those requiring instance-level correspondence. For consistency, we introduce localized edits in shared and view-specific regions, testing whether models distinguish genuine scene contradictions from differences caused by viewpoint and visibility. For spatial reasoning, we construct object relations that require combining spatial evidence across viewpoints to compare physical proximity. We evaluate twelve state-of-the-art open-weight and proprietary VLMs and find that counting accuracy drops substantially on complementary views, ranging from 4.5–50.0%, compared with 45.5–81.5% on complete-overlap examples. Moreover, cross-view consistency remains around 50% across models, including recent proprietary VLMs, despite substantially stronger performance on spatial reasoning. Our results reveal a substantial gap between processing multiple images and integrating evidence across them, highlighting the need for evaluations that explicitly test scene-level multi-view understanding. Code and data are available at <https://github.com/swathisuhas/OneScene-Bench>.

## 1 Introduction

Vision-language models (VLMs) increasingly support multiple images as input, enabling them to reason over different observations of the same physical scene. Unlike generic multi-image understanding, multi-view scene understanding requires models to recognize repeated objects across viewpoints, account for viewpoint-dependent visibility, and combine complementary observations into a coherent representation of the scene. These capabilities are particularly important in embodied settings such as robotics and autonomous systems, where an agent observes the same environment from multiple viewpoints but must reason about it as a single scene.

However, multi-view input does not necessarily imply multi-view reasoning. Consider counting unique objects across two views. A model should recognize repeated instances across viewpoints and count each physical object only once. Yet when both views contain the same objects, or one contains only a subset of the other, taking the larger per-view count gives the correct answer. When each view contains instances not visible in the other, the model must instead determine which objects are shared to recover the total. A model with scene-level understanding should correctly integrate



Figure 1: **Overview of OneScene-Bench.** OneScene-Bench evaluates whether VLMs correctly integrate and reason across views of the same physical scene through three tasks: unique object counting under complete-overlap, subset-overlap, and complementary-view configurations; cross-view consistency; and spatial reasoning. Representative image pairs, questions, and ground-truth answers are shown for each task.

information across different configurations, regardless of how objects are distributed across the views. This reveals a fundamental ambiguity in multi-view evaluation: a correct answer does not necessarily show cross-view integration. Recent benchmarks evaluate reasoning across multiple images and viewpoints in diverse environments [8, 18, 37, 42]. However, these evaluations do not explicitly distinguish between predictions that require cross-view integration and those that can be solved through independent per-view processing. As a result, strong benchmark performance does not necessarily mean that a model has integrated and correctly linked information across views. A reliable evaluation must therefore control whether cross-view integration is actually necessary to obtain the correct answer.

We introduce **OneScene-Bench**, a human-annotated, controlled benchmark for evaluating whether VLMs reconcile multiple views as observations of the same physical scene (Figure 1). Our central design principle is to control the evidence required to solve each example. For multi-view counting, we construct three configurations of cross-view object overlap: *complete overlap*, *subset overlap*, and *complementary views*. The first two admit a max-count shortcut, whereas in complementary views each image contributes instances not visible in the other, making per-view counts alone insufficient. Comparing performance across these configurations directly measures whether counting ability persists when cross-view integration becomes necessary. We probe the same capability with a second task, **cross-view consistency**. Differences between two views do not necessarily imply differences in the scene, as objects may be out of frame or occluded. We introduce localized edits either in regions visible from both viewpoints, creating a genuine contradiction between the views, or in regions visible from only one, leaving the pair compatible. This tests whether a model understands which evidence should agree across viewpoints, rather than merely detecting differences between images. Finally, **multi-view spatial reasoning** evaluates whether models can combine evidence across viewpoints to compare the physical proximity of objects. All examples are constructed from real-world multi-view captures [21] and manually verified.

We evaluate twelve open-weight and proprietary VLMs using an open-ended protocol and analyze both their performance and failure modes. Across all models, counting performance drops on complementary views: accuracy ranges from 45.5–81.5% on complete overlap and 26.7–76.7% on subset overlap, but only 4.5–50.0% on complementary views, where cross-view correspondence and deduplication are required. Cross-view consistency accuracy remains around 50%, i.e., random

chance, across models, with models responding similarly to edits in shared and view-specific regions, indicating limited sensitivity to cross-view visibility. We further attribute counting predictions to simple per-view aggregation strategies and study sensitivity to image order, removal of visual input, and explicit reasoning. Together, these results reveal a clear gap between processing multiple images and correctly performing visual reasoning across them.

We summarize our contributions as follows:

- We introduce **OneScene-Bench**, comprising 1,300 manually verified examples from real-world multi-view scenes across counting, consistency, and spatial reasoning.
- We develop **controlled diagnostics for cross-view integration**: an overlap-based counting taxonomy that distinguishes cases where per-view aggregation is sufficient from those where cross-view integration is necessary, and a consistency intervention that controls whether visual differences occur in regions shared across viewpoints.
- We provide a broad evaluation of **twelve VLMs** under an open-ended protocol, together with analyses of aggregation behavior, image-order sensitivity, no-image bias, explicit reasoning, and scaling with additional views, revealing a persistent gap between processing multiple images and integrating them into a coherent scene-level representation.

## 2 Related Work

**Multimodal Large Language Models (MLLMs).** Early MLLMs established the main paradigms for integrating vision with large language models. Flamingo [2] introduced gated cross-attention to condition an LLM on interleaved images and text, enabling multimodal few-shot learning. BLIP-2 [20] proposed the Q-Former as an efficient interface between frozen vision encoders and LLMs, while LLaVA [22] demonstrated the effectiveness of visual instruction tuning for general visual conversation and reasoning. Building on these foundations, proprietary models such as GPT-4V [26], Gemini [29, 30], and Claude 3 [4] demonstrated broad multimodal reasoning across images, videos, and other modalities. Similarly, open-weight models, including the Qwen-VL series [6, 34, 7], LLaVA-OneVision [19], InternVL [35], MiniCPM-V [39], GLM-4.6V [16], Phi-4-Multimodal [1], and Kimi-VL [32], extended visual instruction tuning to high-resolution images, multiple images, videos, and long multimodal contexts. We evaluate representative models from both proprietary and open-weight families.

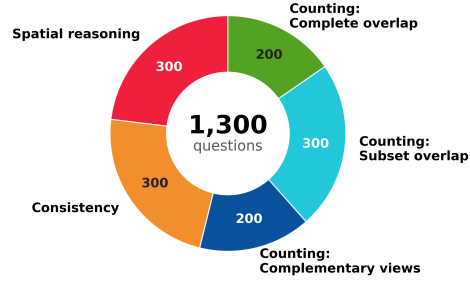
**Multi-Image Understanding Benchmarks.** Multi-image evaluation has progressed from task-specific benchmarks to broader evaluation suites. BLINK [14], Mementos [36], and MileBench [11] focus on visual perception, image sequences, and long multimodal contexts, respectively. MIRB [41], MIBench [23], MMIU [25], MIMIC [10], MUIRBench [33], and MMRB [9] further evaluate diverse perceptual, semantic, and reasoning relations across images. However, these benchmarks often combine related or sequential images rather than different views of the same physical scene. Multi-view benchmarks study this more specific setting across several domains. MV-RoboBench [12] targets spatial understanding and robotic execution, while UrBench [42] and CityCube [37] evaluate urban understanding across ground-level, vehicle, aerial, and satellite views. All-Angles Bench [38] includes counting and geometric reasoning in real-world scenes, whereas MultiView-Bench [40] and TriViewBench [8] study world-centric reasoning and scene recovery in synthetic 3D environments. SpatialMosaic-Bench [18] focuses on integrating fragmented spatial evidence under occlusion and limited overlap. Although these benchmarks cover important aspects of multi-view understanding, they generally do not explicitly separate questions solvable through independent per-view processing from those requiring cross-view integration. OneScene-Bench isolates this distinction by controlling object overlap for counting, the visibility of localized edits for consistency, and the distribution of relational evidence for spatial reasoning. These controlled tasks test whether models reconcile multiple observations into a coherent object-centric representation of the same physical scene.

## 3 OneScene-Bench

We introduce **OneScene-Bench**, a controlled benchmark for evaluating whether vision-language models (VLMs) integrate multiple views into a coherent representation of the same physical scene. Our central design principle is to control whether cross-view integration is necessary to answer a

| Benchmark                     | Multi-view | Free-form responses | Diagnostic | Cross-view consistency | Counting types |
|-------------------------------|------------|---------------------|------------|------------------------|----------------|
| <i>Multi-image benchmarks</i> |            |                     |            |                        |                |
| BLINK                         | ~          | ×                   | ~          | ×                      | ×              |
| MMIU                          | ~          | ×                   | ~          | ×                      | ×              |
| MIRB                          | ×          | ~                   | ~          | ×                      | ×              |
| <i>Multi-view benchmarks</i>  |            |                     |            |                        |                |
| All-Angles Bench              | ✓          | ×                   | ✓          | ~                      | ~              |
| MultiView-Bench               | ✓          | ✓                   | ✓          | ×                      | ×              |
| SpatialMosaic-Bench           | ✓          | ~                   | ✓          | ×                      | ~              |
| OneScene-Bench                | ✓          | ✓                   | ✓          | ✓                      | ✓              |

(a) **Benchmark Properties.** Comparison of representative multi-image and multi-view benchmarks with OneScene-Bench across key benchmark characteristics.



(b) **Benchmark composition.**

Figure 2: **Properties and Composition of OneScene-Bench.** (a) Comparison with representative multi-image and multi-view benchmarks. *Multi-view* denotes multiple observations of the same physical scene; *Free-form* denotes unrestricted responses; *Diagnostic* denotes controlled behavioral analyses; and *Cross-view consistency* denotes explicit consistency evaluation. *Counting types* denotes different counting scenarios covering different levels of view overlap. Green ticks, orange tildes, and red crosses indicate full, partial, and no coverage, respectively. (b) Composition of the 1,300-question core two-view benchmark. Multi-view counting contains 700 questions: 200 complete-overlap, 300 subset-overlap, and 200 complementary-view examples. Cross-view consistency and multi-view spatial reasoning contain 300 questions each.

question correctly. This distinction is important because some multi-view questions can be solved by processing each image independently and applying a simple aggregation rule, without establishing correspondence between observations. OneScene-Bench evaluates three complementary capabilities: **multi-view counting**, **cross-view consistency**, and **multi-view spatial reasoning**.

The benchmark is constructed from DL3DV-10K [21], which contains real-world scenes captured from multiple viewpoints. The core benchmark comprises **1,300 human-verified questions**, with two views per example: 700 counting questions, 300 consistency questions, and 300 spatial-reasoning questions (Fig. 2b). The required outputs are an integer count, a binary YES/NO judgment, and a choice between two reference objects, respectively. We use model-assisted candidate mining (see Appendix A) to identify suitable examples at scale, but all questions follow fixed human-defined templates and all annotations are manually verified.

### 3.1 Multi-View Counting

Given two views of the same scene and a target object category, the model must predict the number of *unique physical instances* visible across the views. Let  $S_1$  and  $S_2$  denote the sets of target instances visible in each view. The scene-level count is

$$C = |S_1 \cup S_2|. \quad (1)$$

We control the relationship between  $S_1$  and  $S_2$  to distinguish cases that admit a simple per-view shortcut from those that require cross-view correspondence.

**Complete overlap.** Both views contain the same target instances,  $S_1 = S_2$ . The correct scene-level count equals either per-view count. We include 200 examples.

**Subset overlap.** One view contains a subset of the instances in the other,  $S_1 \subset S_2$  or  $S_2 \subset S_1$ . The correct answer can still be recovered as the maximum of the two per-view counts. We include 300 examples.

**Complementary views.** Each view contains at least one instance absent from the other,  $S_1 \setminus S_2 \neq \emptyset$  and  $S_2 \setminus S_1 \neq \emptyset$ . Neither view alone contains all instances, and the maximum per-view count is insufficient. Correct prediction therefore requires identifying repeated instances across views while incorporating instances visible from only one viewpoint. We include 200 examples.

**Solvability.** All three configurations require aggregating observations across views to recover a scene-level count, but they differ in the cross-view information required for this aggregation. In complete overlap, both views contain the same set of physical instances, while in subset overlap, the instances in one view are contained in the other. In both cases, the union in Eq. (1) equals the larger per-view set, so the scene-level count can be recovered without explicitly estimating instance overlap. In complementary views, neither view contains the complete set of instances; recovering the union additionally requires identifying which observations correspond to the same physical instances across views, i.e., determining  $|S_1 \cap S_2|$ . These configurations therefore preserve the need for multi-view aggregation while controlling whether explicit instance-level correspondence is required.

### 3.2 Cross-View Consistency

Given two views, the model determines whether they are compatible as observations of the *same underlying scene*. This task requires reasoning about viewpoint-dependent visibility: an object appearing in only one image does not necessarily indicate a contradiction because its location may be outside the field of view or occluded in the other image. We construct controlled examples by introducing a localized object or visual artifact into one view using gpt-image-1.5 [27]. We vary only whether the edited location should be observable from the other viewpoint. An edit in a *view-specific region*, which is not observable from the other view, leaves the pair compatible with the same scene; the pair is labeled CONSISTENT. An edit in a *shared observable region* creates a contradiction between the two observations; the pair is labeled INCONSISTENT. Both conditions therefore contain an image-level difference, while differing in whether that difference implies a scene-level contradiction; detecting differences between images alone cannot separate them. We construct 300 balanced examples using this procedure.

### 3.3 Multi-View Spatial Reasoning

The spatial-reasoning task evaluates whether VLMs can integrate spatial evidence across viewpoints. Each question specifies three objects,  $A$ ,  $B$ , and  $C$ , and asks whether  $B$  is physically closer to  $A$  or to  $C$ . Notably, these objects might be spread over both views, for example, with  $A$  and  $B$  visible in one image and  $B$  and  $C$  in the other. We construct 300 questions from pairs of views of the same DL3DV-10K [21] scene. Qwen3-VL-32B-Instruct [7] proposes candidate object triplets for which the two views provide complementary evidence about relative proximity. We favor examples in which combining evidence across viewpoints informs the comparison beyond what is readily available from either view alone. Selected triplets are instantiated using a fixed question template and manually verified.

### 3.4 Multi-View Scaling

The core benchmark uses two views per example. To study how counting behaves as observations accumulate, we extend the 200 complete-overlap questions to  $N \in \{2, \dots, 6\}$  input views, keeping all target instances visible in every selected view so that the scene-level count is unchanged. Each additional view contributes another observation of the same physical instances while increasing the number of cross-view correspondences the model must resolve. This extension therefore measures whether additional viewpoints improve object recognition and scene understanding or instead increase identity confusion and double-counting. Evaluating every question in all five settings yields  $200 \times 5 = 1,000$  view-conditioned instances; the two-view versions already belong to the core benchmark, so the extension adds 800 configurations, for 2,100 evaluation configurations in total.

**Human Verification.** All automatically proposed or generated examples undergo manual review before inclusion in OneScene-Bench; the VLM and image-generation model serve only to accelerate candidate discovery and controlled editing, and their outputs are never treated directly as ground truth. For counting, annotators verify the target category, the number of visible instances in each image, the identity of repeated instances across views, the overlap type, and the unique scene-level count. For consistency, annotators verify whether the edited region is shared between the viewpoints or visible from only one view, and whether the resulting consistency label is correct. For spatial reasoning, annotators confirm the identities of the three objects and the relative-proximity answer. Incorrect annotations are corrected, while examples with uncertain object identity, visibility, correspondence, or spatial relationships are removed.

| Model  | Counting $\uparrow$               |             |               | Consistency $\uparrow$ | Spatial Reasoning $\uparrow$ |
|--------|-----------------------------------|-------------|---------------|------------------------|------------------------------|
|        | Complete                          | Subset      | Complementary |                        |                              |
| Open   | Qwen3-VL-32B-Instruct [7]         | 70.0        | <b>70.7</b>   | 34.0                   | 49.3                         |
|        | Gemma-4-31B-IT [31]               | 61.0        | 63.3          | <b>35.0</b>            | 50.3                         |
|        | InternVL3.5-38B-HF [35]           | <b>72.0</b> | 57.3          | 14.5                   | 50.7                         |
|        | LLaVA-OneVision-2-8B-Instruct [3] | 55.5        | 53.0          | 18.0                   | 50.0                         |
|        | MiniCPM-V-4.5 [39]                | 67.0        | 51.7          | 20.5                   | 53.3                         |
|        | Phi-4-multimodal-instruct [1]     | 54.0        | 26.7          | 4.5                    | 50.0                         |
|        | GLM-4.6V-Flash [16]               | 57.0        | 56.7          | 28.0                   | <b>55.7</b>                  |
|        | Kimi-VL-A3B-Instruct [32]         | 66.0        | 46.3          | 10.0                   | 50.0                         |
| Closed | GPT-4o [17]                       | 61.5        | 52.3          | 28.5                   | 47.0                         |
|        | GPT-5.5 [28]                      | 70.0        | 69.3          | 38.5                   | 54.3                         |
|        | Gemini 3.7 Flash [15]             | <b>81.5</b> | <b>76.7</b>   | <b>50.0</b>            | 49.0                         |
|        | Claude Haiku 4.5 [5]              | 45.5        | 43.3          | 18.5                   | <b>55.3</b>                  |

Table 1: **VLMs on OneScene-Bench.** Accuracy (%) of twelve open-weight and proprietary VLMs on OneScene-Bench. The best-performing model in each group is shown in bold for every task. Across models, counting performance drops substantially on complementary views, where object correspondence and deduplication across viewpoints are required. Cross-view consistency remains around 50%, whereas spatial reasoning is substantially stronger, revealing a marked difference in model capability across forms of multi-view integration.

## 4 Evaluation

### 4.1 Evaluation Setup

We evaluate all models in five settings in OneScene-Bench: complete-overlap counting, subset-overlap counting, complementary-view counting, cross-view consistency, and multi-view spatial reasoning. All models receive the same task instructions and input images, and we retain their complete generated responses.

We use GPT-4o [17] as an LLM judge for all benchmark results. The judge receives the question, ground-truth answer, and complete model response, and determines whether the response reaches the correct conclusion. We report judge-based accuracy, allowing semantically equivalent answers and minor differences in response format. The exact prompts used are available in Appendix C.

### 4.2 Benchmark Evaluation Results

Table 1 reports accuracy across the five evaluation settings. Qwen3-VL-32B-Instruct [7] drops from 70.0% and 70.7% on complete- and subset-overlap examples to 34.0% on complementary views, while InternVL3.5-38B-HF [35] drops from 72.0% to 14.5%. Even the strongest models exhibit this pattern: Gemini 3.7 Flash [15] decreases from 81.5% to 50.0%, while GPT-5.5 [28] drops from 70.0% to 38.5%. The consistent degradation on complementary views identifies cross-view correspondence and deduplication as a key bottleneck, where models perform substantially better when the views contain largely redundant evidence than when information must be reconciled across views.

Consistency accuracy remains close to chance, ranging from 47.0% to 55.7%. LLaVA-OneVision-2-8B-Instruct [3] and Phi-4-multimodal-instruct [1] predict YES for all examples and consequently obtain 50.0% on the balanced dataset. This behavior suggests that models often default to treating different views as consistent rather than verifying whether scene content agrees across viewpoints.

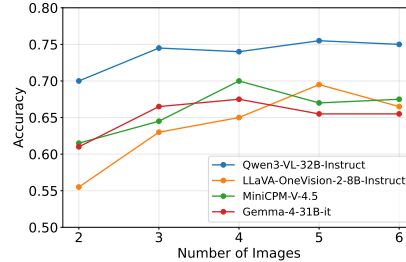
Spatial reasoning capabilities are substantially stronger for most models. GPT-5.5 [28] achieves the highest accuracy at 92.7%, followed closely by Gemini 3.7 Flash [15] at 89.7%. Kimi-VL-A3B-Instruct is a notable outlier at 15.7%, frequently selecting object *B* rather than one of the required reference objects *A* or *C*, indicating a failure to follow the specified answer format. Together, these results show that current VLMs can perform well on some multi-view tasks, yet struggle when they must consistently reconcile information across viewpoints.

| Model                             | Counting            |                   |                          | Consistency $\uparrow$ | Spatial Reasoning $\uparrow$ |
|-----------------------------------|---------------------|-------------------|--------------------------|------------------------|------------------------------|
|                                   | Complete $\uparrow$ | Subset $\uparrow$ | Complementary $\uparrow$ |                        |                              |
| Qwen3-VL-32B-Instruct [7]         | 37.5 (70.0)         | 26.7 (70.7)       | 7.0 (34.0)               | 50.0 (49.3)            | 63.7 (85.3)                  |
| Gemma-4-31B-IT [31]               | 34.0 (61.0)         | 28.3 (63.3)       | 13.5 (35.0)              | 0.0 (50.3)             | 1.7 (86.3)                   |
| InternVL3.5-38B-HF [35]           | 41.0 (72.0)         | 14.3 (57.3)       | 1.5 (14.5)               | 50.0 (50.7)            | 55.3 (80.7)                  |
| LLaVA-OneVision-2-8B-Instruct [3] | 40.0 (55.5)         | 17.7 (53.0)       | 3.5 (18.0)               | 50.0 (50.0)            | 54.7 (77.0)                  |

Table 2: **No-image bias.** Accuracy (%) when models receive the task instructions but no images; performance with images is shown in gray in parentheses. Models achieve non-trivial accuracy without visual input, indicating response biases and task priors, while visual evidence substantially improves counting and spatial reasoning performance.

| Model                         | Counting Accuracy   |                   |                          |
|-------------------------------|---------------------|-------------------|--------------------------|
|                               | Complete $\uparrow$ | Subset $\uparrow$ | Complementary $\uparrow$ |
| Qwen3-VL-32B-Instruct [7]     | <b>70.0</b>         | <b>70.7</b>       | <b>34.0</b>              |
| Qwen3-VL-32B-Thinking [7]     | 59.0                | 57.0              | 26.5                     |
| Kimi-VL-A3B-Instruct [32]     | <b>66.0</b>         | <b>46.3</b>       | 10.0                     |
| Kimi-VL-A3B-Thinking [32]     | 21.5                | 38.0              | <b>11.5</b>              |
| MiniCPM-V-4.5 [39]            | <b>67.0</b>         | <b>51.7</b>       | <b>20.5</b>              |
| MiniCPM-V-4.5 (Thinking) [39] | 61.5                | 44.3              | 9.5                      |

(a) Reasoning-enabled variants.



(b) Number of input views.

Figure 3: **Diagnostic analyses of reasoning and number of views.** (a) Counting accuracy of instruction-tuned and reasoning-enabled variants. Reasoning generally reduces performance, indicating that explicit reasoning does not improve multi-view counting. (b) Complete-overlap counting accuracy with  $N = 2$ –6 input views. Performance initially improves with additional views but peaks before  $N = 6$  for all models, indicating that more views do not yield monotonic gains.

### 4.3 Diagnostic Analyses

We conduct additional analyses to examine whether model behavior is affected by input ordering, reliance on non-visual priors, explicit reasoning, simple per-view aggregation, and the number of input views.

**No-Image Bias.** We remove all input images while keeping the questions and evaluation procedure unchanged. Although accuracy generally drops, models retain non-trivial performance on several tasks, revealing substantial language and answer-distribution priors. Notably, LLaVA-OneVision-2 obtains 50% consistency accuracy both with and without images by predicting YES for every example, showing that unchanged performance need not indicate successful use of visual evidence. Full results are reported in Table 2; Appendix B.2 lists how often models refuse to answer because no image is provided.

**Instruct vs. Thinking Variants.** We compare matched reasoning-enabled and instruction-tuned variants of Qwen3-VL [7] and Kimi-VL [32], together with direct and thinking-enabled inference for MiniCPM-V-4.5 [39]. As shown in Fig. 3a, reasoning reduces counting accuracy across nearly all models and settings; the only exception is a 1.5-point gain for Kimi-VL on complementary views. Thus, explicit reasoning does not systematically improve multi-view counting.

**Comparison with Aggregation Heuristics.** We test whether joint predictions can be explained by simple aggregation of independent per-view counts. Let  $C_1$  and  $C_2$  denote individual-view predictions and  $C_3$  the joint prediction; we report the percentage of questions for which  $C_3$  agrees with either  $C_1 + C_2$  or  $\max(C_1, C_2)$ . Across models, joint predictions agree substantially more often with the maximum than with the sum (Table 3). In complete- and subset-overlap settings, maximum aggregation can recover the scene-level count when per-view counts are correct. In complementary views, unique instances are distributed across views, requiring information from both images. Yet maximum agreement remains consistently higher than sum agreement, despite maximum aggregation being insufficient to recover the correct scene-level count.

| Model                             | Complete overlap |          |          | Subset overlap |          |          | Complementary views |          |          |
|-----------------------------------|------------------|----------|----------|----------------|----------|----------|---------------------|----------|----------|
|                                   | Acc. ↑           | Sum agr. | Max agr. | Acc. ↑         | Sum agr. | Max agr. | Acc. ↑              | Sum agr. | Max agr. |
| Qwen3-VL-32B-Instruct [7]         | 70.0             | 8.0      | 82.0     | 70.7           | 23.0     | 77.7     | 34.0                | 18.0     | 39.0     |
| InternVL3.5-38B-HF [35]           | 72.0             | 4.0      | 77.0     | 57.3           | 15.0     | 75.3     | 14.5                | 11.0     | 51.0     |
| LLaVA-OneVision-2-8B-Instruct [3] | 55.5             | 13.5     | 68.0     | 53.0           | 10.3     | 64.3     | 18.0                | 4.5      | 62.0     |
| Gemma-4-31B-IT [31]               | 61.0             | 8.5      | 69.5     | 63.3           | 21.7     | 70.3     | 35.0                | 17.5     | 34.0     |

Table 3: **Aggregation Heuristics.** Comparison of joint and single-view counting predictions.  $C_1$  and  $C_2$  denote individual-view predictions and  $C_3$  the joint prediction. *Acc.* is ground-truth accuracy, shown in gray for reference. *Sum agr.* and *Max agr.* report the percentage of questions satisfying  $C_3 = C_1 + C_2$  and  $C_3 = \max(C_1, C_2)$ , respectively, without using ground truth. Joint predictions often agree with the maximum in complete- and subset-overlap settings; in complementary views, maximum agreement remains higher than sum agreement despite being insufficient for scene-level counting.

**Effect of Number of Views.** We evaluate complete-overlap counting with  $N \in \{2, 3, 4, 5, 6\}$  views while keeping the scene-level count fixed. Figure 3b shows that performance initially improves with additional views but peaks before  $N = 6$  for all four models, suggesting that additional visual evidence is useful only up to the point where integrating more views becomes increasingly difficult. Moreover, there is still a substantial gap in accuracy that none of the models can close, suggesting there is plenty of room for improvement in visual reasoning capabilities when considering scenes given as multiple views.

## 5 Discussion

Our results reveal a clear gap between processing multiple images and reasoning over them coherently in the current generation of VLMs. This is most evident in complementary-view counting, where performance drops substantially when models must establish cross-view correspondence and deduplicate repeated instances. Cross-view consistency also remains difficult, including for recent models, suggesting that robust scene-level multi-view understanding remains unresolved. The additional diagnostic analyses further expose limitations in how models use and integrate evidence across views.

These findings complement existing multi-view benchmarks, on which models have demonstrated stronger spatial and geometric reasoning. OneScene-Bench shows that strong spatial reasoning does not necessarily imply reliable scene-level integration. With multi-view images becoming a dominant approach for interacting with visual 3D scenes through language, OneScene-Bench offers a new perspective on what is missing for 3D reasoning through powerful LLMs. Current models can reason about spatial relationships while still struggling with object correspondence, deduplication, and cross-view consistency, which are essential for reliable and accurate 3D reasoning.

**Limitations.** OneScene-Bench focuses on three scene-level capabilities and 1,300 manually verified examples from DL3DV-10K. While larger datasets and broader model coverage would be valuable, we prioritize careful human verification and balance evaluation breadth against computational and API costs. Future work could further expand the benchmark across tasks, environments, and models.

## 6 Conclusion

We introduced OneScene-Bench, a controlled benchmark for evaluating whether vision-language models are able to integrate multiple views of a scene to correctly reason about visual aspects of the scene. Across open-weight and proprietary VLMs, our results revealed a persistent gap between processing multiple images and integrating them: models struggle particularly when they must establish object correspondence, reconcile complementary evidence, and detect inconsistencies across viewpoints. OneScene-Bench provides a controlled framework for measuring these capabilities and tracking progress toward more reliable scene-level multi-view understanding.

This work thus showed that models of different architectures, both closed and open, are usually unable to perform basic visual reasoning across multiple views, the prevalent representation format for 3D scenes. Closing this gap becomes essential when moving to more complex 3D reasoning with VLMs, and our OneScene-Bench is an open benchmark designed to track progress toward this goal.

## References

- [1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [3] Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.
- [4] Anthropic. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic, 2024.
- [5] Anthropic. Claude haiku 4.5 system card. Technical report, Anthropic, October 2025.
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.
- [7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- [8] Yu-Yang Chen and Lan-Zhe Guo. Triviewbench: Controlled complexity scaling for multi-view structural reasoning in mllms. *arXiv preprint arXiv:2606.26029*, 2026.
- [9] Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, et al. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*, 2025.
- [10] Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More images, more problems? a controlled analysis of vlm failure modes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 7423–7442, 2026.
- [11] Song Dingjie, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. In *First Conference on Language Modeling*, volume 2, page 3, 2024.
- [12] Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, et al. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. *arXiv preprint arXiv:2510.19400*, 2025.
- [13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [14] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- [15] Google DeepMind. Gemini 3.7 flash model card. Technical report, Google DeepMind, August 2026.
- [16] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025.

- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [18] Kanghee Lee, Injae Lee, Minseok Kwak, Jungi Hong, Kwonyoung Ryu, and Jaesik Park. Spatialmosaic: A multiview vlm dataset for partial visibility. *arXiv preprint arXiv:2512.23365*, 2025.
- [19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PmLR, 2023.
- [21] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22160–22169. IEEE, 2024.
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [23] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428, 2024.
- [24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [25] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 38405–38453, 2025.
- [26] OpenAI. GPT-4V(ision) system card. Technical report, OpenAI, September 2023.
- [27] OpenAI. The new ChatGPT images is here. OpenAI Blog, December 2025.
- [28] OpenAI. GPT-5.5 System Card. <https://openai.com/index/gpt-5-5-system-card/>, April 2026. Accessed: 2026-08-31.
- [29] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [30] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [31] Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.
- [32] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [33] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pages 62624–62650, 2025.

- [34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [35] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.
- [36] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442, 2024.
- [37] Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyu Wang, Junren Rao, Wenhao Lu, Weishi Li, Qunjun Yin, and Yong Li. Citycube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9190–9215, 2026.
- [38] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12000–12008, 2026.
- [39] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11704–11715, 2026.
- [40] Hantao Zhang, Jinru Sui, Ed Li, Dirk Bergemann, and Zhuoran Yang. Multiview-bench: A diagnostic benchmark for world-centric multi-view integration in vlms. *arXiv preprint arXiv:2607.08970*, 2026.
- [41] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.
- [42] Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. 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, pages 10707–10715, 2025.

## A Detailed Dataset Construction

This section provides additional details on the construction and verification procedure used to build OneScene-Bench.

### A.1 Counting Candidate Mining

For counting, we first randomly select a location from DL3DV-10K [21] and randomly sample two frames from its capture sequence. We require the difference between their frame indices to be between 10 and 25, providing viewpoint variation while retaining sufficient visual overlap. The selected pair is provided to Qwen3-VL-32B-Instruct [7], which is asked to identify a clearly visible, countable, and unambiguous object category. For each proposed category, the model estimates the number of visible instances in each view and the number of unique physical instances across the two views.

If no suitable object category is identified, another pair is sampled from the same location and the process is repeated. We allow up to 10 attempts per location; locations for which no suitable candidate is found are skipped. This procedure is repeated across locations until sufficient candidates are obtained for the three counting configurations.

The final overlap type is assigned manually after verifying the visible instance sets. The resulting candidates are divided into complete-overlap, subset-overlap, and complementary-view examples, yielding 200, 300, and 200 questions, respectively.

### A.2 Consistency Construction

For the consistency task, we again begin by selecting two frames from the same DL3DV-10K [21] location using the 10–25 frame-index separation. To identify shared and view-specific regions, we first establish local feature correspondences between the two views using SIFT [24]. We use RANSAC [13] to remove geometrically inconsistent matches. Regions containing a dense concentration of verified correspondences are treated as shared across the two views, while regions spatially separated from these correspondences and containing few matched features are selected as candidate view-specific regions. Image pairs for which these regions cannot be identified reliably are discarded.

We then introduce a localized object or visual artifact into one of these regions using gpt-image-1.5 [27]. The model is given freedom to select an appropriate object to insert, while the target region is determined by the desired consistency condition. An edit in a view-specific region is not expected to be observed from the second viewpoint, so the resulting pair is labeled *consistent*. In contrast, an edit in a shared region introduces a change that should also be observable from the second viewpoint; because it appears in only one view, the resulting pair is labeled *inconsistent*.

The edited patch is resized to the target region and blended into the original image using a Gaussian-smoothed mask to reduce boundary artifacts. All generated examples are subsequently manually inspected to verify the edit, its visibility across the two viewpoints, and the resulting consistency label. Ambiguous or invalid examples are rejected or corrected. We construct 300 balanced consistency examples.

### A.3 Spatial Reasoning Candidate Mining

For spatial reasoning, we again sample two frames from the same DL3DV-10K [21] location with a frame-index separation of 10–25. Qwen3-VL-32B-Instruct [7] is used to identify three clearly visible objects that provide a meaningful relative-proximity comparison across the two views. The selected objects are denoted by  $A$ ,  $B$ , and  $C$ , and the candidate question asks whether  $B$  is physically closer to  $A$  or to  $C$ .

A dedicated annotation tool is then used to assign consistent object names and record the ground-truth proximity relationship. This allows the object references and the final answer to be standardized independently of the wording proposed during candidate generation. Candidates are retained only when the three objects can be reliably identified across the views and the relative-proximity relationship is unambiguous. We construct 300 spatial reasoning questions.

| Model                             | Counting   |          |                 | Consistency ↓ | Spatial Reasoning ↓ |
|-----------------------------------|------------|----------|-----------------|---------------|---------------------|
|                                   | Complete ↓ | Subset ↓ | Complementary ↓ |               |                     |
| Qwen3-VL-32B-Instruct [7]         | 0.5        | 2.4      | 1.5             | 4.0           | 2.7                 |
| Gemma-4-31B-IT [31]               | 6.0        | 1.6      | 4.0             | 0.3           | 0.3                 |
| InternVL3.5-38B-HF [35]           | 3.5        | 1.0      | 1.5             | 2.3           | 1.0                 |
| LLaVA-OneVision-2-8B-Instruct [3] | 1.5        | 0.3      | 0.5             | 0.0           | 9.0                 |

Table 4: **Image-Order Sensitivity.** Absolute difference in accuracy (percentage points) between the original and reversed input-image orders,  $|\text{Acc}_{\text{original}} - \text{Acc}_{\text{reverse}}|$ . Overall, 18 of 20 model–task combinations change by at most four points, indicating broad robustness to image order, with larger sensitivity for Gemma-4 on complete-overlap counting and LLaVA-OneVision-2 on spatial reasoning.

#### A.4 Verification Procedure

All candidate examples are manually verified before inclusion in the benchmark. Reviewers inspect the selected images, referenced objects, and ground-truth annotations and reject or correct candidates with ambiguous object identity, uncertain cross-view correspondence, unclear visibility, or ambiguous answers.

For counting, reviewers verify the target category, visible instances in each view, cross-view object correspondences, unique scene-level count, and manually assigned overlap type. For consistency, reviewers verify the edited region, whether it is shared or view-specific, and whether the resulting pair is consistent with a single underlying scene. For spatial reasoning, reviewers verify the three referenced objects and the relative-proximity relationship.

The dedicated interfaces used for counting, consistency, and spatial reasoning verification are shown in Figures 4, 5, and 6, respectively.

## B Additional Details on Evaluation Tasks

### B.1 Image Order Sensitivity

We study whether model predictions depend on the order in which the two views are presented. A model that forms a consistent scene representation should produce similar predictions when the two views are swapped. We evaluate four representative open-weight models on the same questions using both the original and reversed image orders, and measure the absolute change in accuracy across all five evaluation settings. Most model–task combinations are stable, with 18 of 20 changing by at most four percentage points. The largest changes occur for Gemma-4 [31] on complete-overlap counting and LLaVA-OneVision-2-8B-Instruct [3] on spatial reasoning, with differences of 6.0 and 9.0 points, respectively. Overall, models are reasonably robust to image ordering, suggesting that input order is not a major source of variation in the benchmark.

### B.2 No-Image Evaluation Setting

To further characterize model behavior when visual evidence is unavailable, we record whether models explicitly refuse to answer in the no-image setting. A response is counted as a refusal only when the model explicitly states that the question cannot be answered without the missing visual information; incorrect or guessed answers are not counted as refusals. As shown in Table 5, Qwen3-VL-32B-Instruct [7], InternVL3.5-38B-HF [35], and LLaVA-OneVision-2-8B-Instruct [3] do not refuse any questions. In contrast, Gemma-4-31B-IT [31] refuses all consistency questions and 293 of 300 spatial reasoning questions, while refusing only 21 of 700 counting questions. This difference shows that models vary substantially in whether they recognize when visual information is necessary to answer a question.

| Model                             | Counting                  |                         |                                | Consistency   | Spatial                    |
|-----------------------------------|---------------------------|-------------------------|--------------------------------|---------------|----------------------------|
|                                   | Complete<br>( $n = 200$ ) | Subset<br>( $n = 300$ ) | Complementary<br>( $n = 200$ ) | ( $n = 300$ ) | Reasoning<br>( $n = 300$ ) |
| Qwen3-VL-32B-Instruct [7]         | 0                         | 0                       | 0                              | 0             | 0                          |
| InternVL3.5-38B-HF [35]           | 0                         | 0                       | 0                              | 0             | 0                          |
| LLaVA-OneVision-2-8B-Instruct [3] | 0                         | 0                       | 0                              | 0             | 0                          |
| Gemma-4-31B-IT [31]               | 6                         | 3                       | 12                             | 300           | 293                        |

Table 5: **No-image refusal count.** Number of explicit refusals in the no-image setting. A refusal is counted only when a model explicitly states that the question cannot be answered without visual input; incorrect or guessed answers are not counted.

## C Evaluation Prompts

For reproducibility, we provide the exact prompts used for model evaluation. The same task prompts are used for all evaluated models. The free-form model responses are evaluated using an LLM judge, which receives the raw model response and the human-verified ground-truth answer.

### C.1 Task Prompts

#### Multi-View Counting.

These two images show the same scene from different viewpoints. How many distinct instances of '{object\_name}' are visible in the scene? Do not double-count the same instance if it appears in both views. Use both images jointly to infer the scene-level count. Reply with only one integer and no explanation.

Here, {object\_name} is replaced by the target object category for each example. The same prompt is used for all three counting configurations. For the multi-image ablation, we slightly modify the counting prompt to refer to multiple input images rather than two images, while keeping the remaining instructions unchanged.

#### Cross-View Consistency.

These two images are different views of a scene. Check whether they are visually consistent.  
 Answer YES if the images show the same scene contents without any clear contradiction. Answer NO if there is any clear inconsistency, such as a missing object, extra object, changed object, or contradictory scene content.  
 Write exactly: Answer: YES or Answer: NO

#### Multi-View Spatial Reasoning.

These two images show the same scene from different viewpoints.  
 A = a B = b C = c  
 Task: Decide whether B is closer to A or C.  
 Output exactly: Answer: <A or C>

Here, {a}, {b}, and {c} are replaced by the corresponding object names for each example.

### C.2 LLM-as-a-Judge Prompts

#### Counting Judge.

You are evaluating the answer of a vision-language model on a counting task. The model was asked to count the number of distinct instances of an object across multiple views of the same scene.

Object: '{object\_name}'

Human-verified ground-truth count: '{ground\_truth}'

Model's raw response: '{raw\_output}'

Determine whether the model's response gives the same final count as the human-verified ground truth.

Accept semantically equivalent representations of the number. For example: - "3" - "three" - "There are three." - "The answer is 3." should all be treated as the count 3.

Ignore irrelevant wording or explanation as long as the model's final intended count is unambiguous.

If the model gives multiple conflicting counts and no unambiguous final answer, mark it INCORRECT.

Output exactly one word: CORRECT or INCORRECT

Here, {object\_name} denotes the target object category, {ground\_truth} is the human-verified scene-level count, and {raw\_output} is the complete response generated by the evaluated model.

### Consistency Judge.

You are evaluating the answer of a vision-language model on a binary visual consistency task.

The original task asked the model to determine whether two views of a scene are visually consistent.

The human-verified ground-truth answer is: '{ground\_truth}' The model's raw response is: '{raw\_output}'

Determine whether the model's final intended answer agrees with the human-verified ground truth.

Interpret the following as equivalent: YES = SAME = CONSISTENT NO = DIFFERENT = INCONSISTENT

The model may include reasoning, explanation, or additional wording. Ignore such wording as long as the final intended answer is unambiguous.

For example: Ground truth: YES Model response: "The two images appear to show the same scene. Answer: YES" => CORRECT Ground truth: NO Model response: "There is an object missing in the second image, so the scenes are inconsistent." => CORRECT Ground truth: NO Model response: "The images are consistent. Answer: YES" => INCORRECT

If the model gives conflicting answers and there is no unambiguous final answer, mark it INCORRECT.

Output exactly one word: CORRECT or INCORRECT

Here, {ground\_truth} is the human-verified consistency label (YES or NO), and {raw\_output} is the complete response generated by the evaluated model.

### Spatial Reasoning Judge.

You are evaluating the answer of a vision-language model on a spatial reasoning task.

The original task asked the model to determine whether object B is closer to object A or object C.

The objects were:

A = '{object\_a}' B = '{object\_b}' C = '{object\_c}'

The human-verified ground-truth answer is: '{ground\_truth}'

The model's raw response is: '{raw\_output}'

Determine whether the model's final intended answer agrees with the human-verified ground truth.

The answer must correspond to either A or C. Accept semantically equivalent responses.

Examples that mean A: - "A" - "Answer: A" - "B is closer to A." - "Object A." - "The '{object\_b}' is closer to the '{object\_a}'."

Examples that mean C: - "C" - "Answer: C" - "B is closer to C." - "Object C." - "The '{object\_b}' is closer to the '{object\_c}'."

The model may include reasoning, explanation, or additional wording. Ignore such wording as long as the final intended answer is unambiguous.

If the model gives multiple conflicting answers and there is no unambiguous final answer, mark it INCORRECT.

Output exactly one word: CORRECT or INCORRECT

Here, {object\_a}, {object\_b}, and {object\_c} denote the three referenced objects, {ground\_truth} is the human-verified answer (A or C), and {raw\_output} is the complete response generated by the evaluated model.

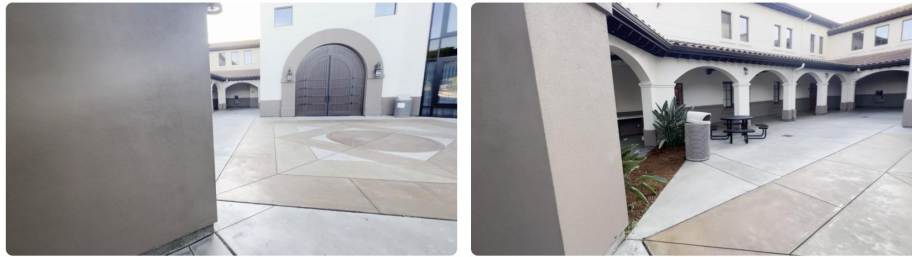
## D Screenshots from the Dataset Verification Tool

Figures 4, 5, and 6 show the interfaces used during manual verification of OneScene-Bench examples for the three benchmark tasks.

### Counting Dataset Review Tool

sample\_000200 | index 199 / 2178

This sample was already reviewed or skipped.



Question:

These are two views of the same location. Count the total number of unique arched doorways present in the location. Do not double-count instances that appear in both images. Answer with a single integer.

Object name / category

arched doorways

Reviewed total count

7

Type

☐ Complete Overlap ☐ Subset Overlap ☒ Complementary View

> What do the types mean?

☐ Skip this image pair

Save reviewed sample

Previous

Next

Figure 4: **Screenshot of the manual counting review interface.** The interface displays the selected multi-view image pair, target object category, reviewed scene-level count, and manually assigned overlap type. Reviewers can also skip ambiguous image pairs before finalizing the annotation.

## Consistency Dataset Review Tool

Sample 303 (302 / 340)

Ground Truth

YES

Label

1

Edit Type

non\_overlap

Object Edited

fire extinguisher

Original Image 1



Final Image 1



Image 2



### Final Check

Everything is correct

✓ Correct — Next Sample

### Ground truth is wrong

Correct ground truth:

☐ yes ☒ no

Update Ground Truth — Next Sample

### Sample itself is wrong

Delete the sample if it is ambiguous, incorrect, or should not be included in the final benchmark.

☐ I confirm that sample 303 should be permanently deleted.

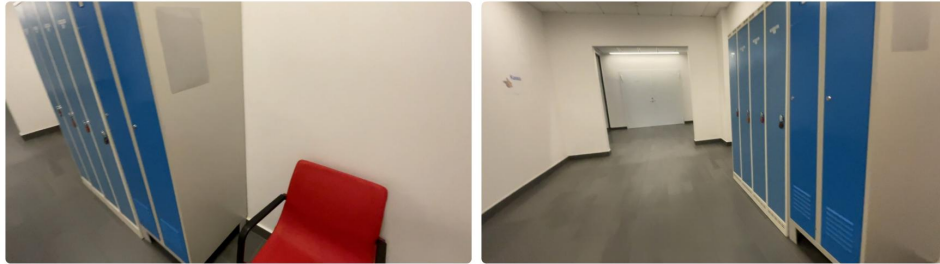
ⓘ Delete Sample

Figure 5: **Screenshot of the consistency review tool.** The interface displays the original and edited views, ground-truth label, edit type, and edited object. Reviewers can confirm the example, correct an incorrect ground-truth label, or permanently remove an invalid or ambiguous example.

## Spatial Reasoning Dataset Review Tool

sample\_000066 | index 65 / 751

This sample was already reviewed or skipped.



Object A

red chair

Object B

blue lockers

Object C

white door

Question:

These are two views of the same location.

A = red chair

B = blue lockers

C = white door

Is B closer to A or C?

Answer with A or C.

Closer to

☒ A ☐ C ☐ unclear

☐ Skip this sample because quality is bad

Save reviewed sample

Previous

Next

Figure 6: **Screenshot of the spatial-reasoning verification interface.** The tool displays the two views, the three referenced objects ( $A$ ,  $B$ , and  $C$ ), the generated question, and the reviewer controls for assigning the ground-truth proximity relationship or marking the example for removal.

## E Qualitative Samples

Figure 7 shows example benchmark questions from all five tasks, along with the answers provided by four different VLMs. The ground-truth answer is marked with a green tick.

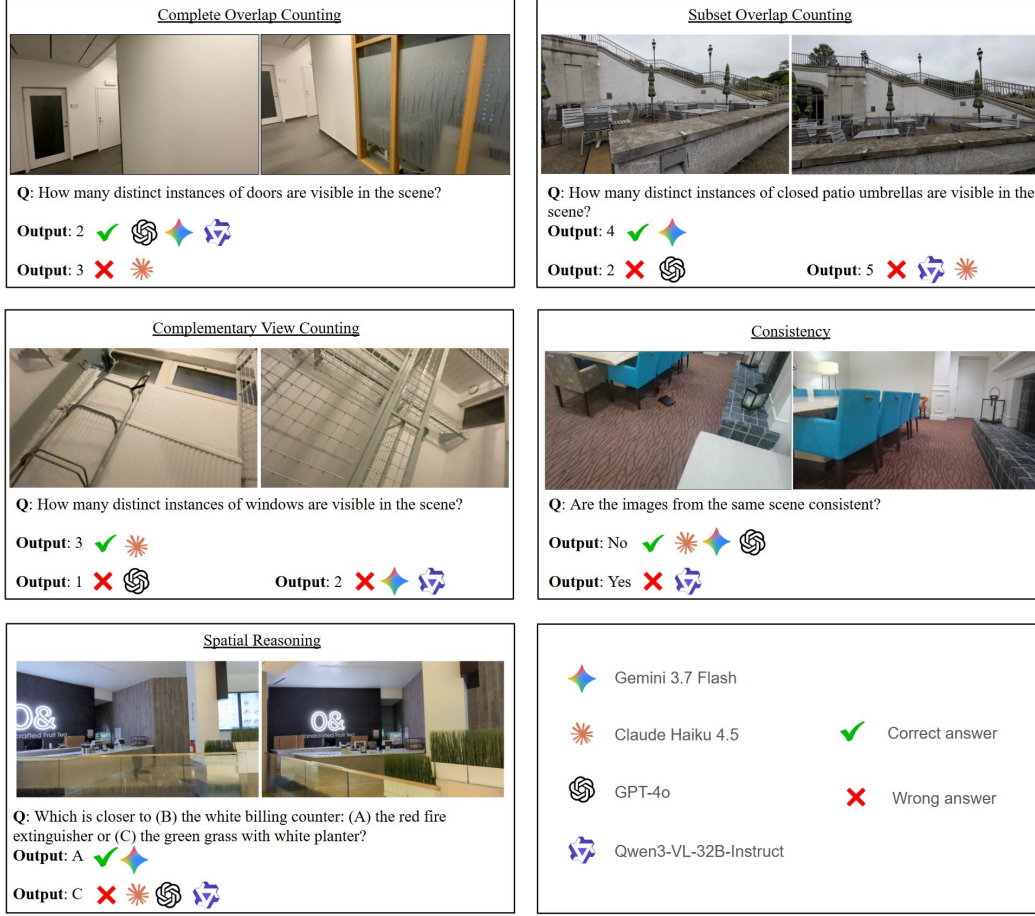


Figure 7: **Representative examples from the five tasks in OneScene-Bench:** multi-view counting under complete-overlap, subset-overlap, and complementary-view configurations, cross-view consistency, and multi-view spatial reasoning. Questions are slightly shortened for visual presentation while preserving their original meaning. Model predictions are shown using the corresponding model logos, with green checkmarks and red crosses indicating correct and incorrect predictions, respectively.

## F Compute Resources

Open-source models were evaluated primarily on NVIDIA A100 GPUs with 80 GB of GPU memory. The closed-source models were evaluated through their respective paid APIs. The evaluation of an open-source model required approximately 2 hours per task.

## G Licenses for Existing Assets

OneScene-Bench is constructed from DL3DV-10K [21], which is used in accordance with the DL3DV-10K Terms of Use. The open-weight models used in our experiments are accessed and used under their respective licenses, including Apache License 2.0 for Qwen3-VL [7], InternVL3.5 [35], LLaVA-OneVision [19, 3], and Gemma 4 [31], and the MIT License for Kimi-VL [32] and GLM-4.6V-Flash [16]. Proprietary models are accessed through their official APIs in accordance with the applicable terms of service.