Fine-Grained Multi-Image Object Hallucination Benchmark

Joonki Min[1*], Chaeyun Kim[1,2*], Hyungwook Choi[1], Yejin Kim[1], Kihyun Kim[1]
Yohan Jo[1†], Joonseok Lee[1†]
[1]Seoul National University [2]AIM Intelligence
{minc69,golddohyun,chooi221,a2000yejin,ki5477,yohan.jo,joonseok}@snu.ac.kr

Abstract

Multimodal Large Language Models (MLLMs) are increasingly deployed in multi-image scenarios requiring complex reasoning across visual contexts. However, current MLLMs remain fundamentally limited by object hallucination—generating plausible yet factually inconsistent descriptions about objects. Existing benchmarks, designed primarily for single-image settings or providing only high-level multi-image assessments, cannot systematically diagnose how visual complexity and reasoning demands trigger hallucination. To address this gap, we introduce MIOH, a fine-grained multi-image object hallucination benchmark that systematically evaluates object hallucination across four foundational tasks (existence, counting, attribute, position) through three multi-image reasoning patterns (comprehensive, comparative, selective) under three controlled adversarial pressures (visual context scale, perceptual difficulty, contextual bias). Through evaluation of 29 models, we reveal that even state-of-the-art systems like GPT-5 and Gemini-2.5-Pro exhibit distinct failure patterns across different reasoning patterns and tasks. Our evaluation reveals that hallucination stems not merely from perceptual failures but from integration-stage limitations when maintaining object representations across multiple images. MIOH provides a controlled framework for analyzing multi-image object hallucination and serves as a critical evaluation tool for developing more reliable multimodal AI systems.

1. Introduction

With recent advances, Multimodal Large Language Models (MLLMs) are capable of reasoning over multiple images simultaneously. This capability requires models not only to recognize content in individual images but also to integrate and synthesize information across diverse visual inputs. Despite these advancements, however, current MLLMs remain fundamentally limited by *object hallucination*, where models generate plausible yet factually inconsistent descriptions about objects in the queried images.

Multi-image scenarios amplify object hallucination along two critical dimensions. First, ***perceptual failure*** : even within individual images, models struggle with visually challenging objects (small size, occlusion, contextual ambiguity) or contextually misleading scenes where co-occurring objects create false expectations. Second, ***information integration*** : multi-image contexts demand integrating and synthesizing information across images, where models must aggregate information across all images **(comprehensive reasoning)**, identify differences between images **(comparative reasoning)**, or retrieve specific images matching given criteria **(selective reasoning)**. Failures in either dimension create compounding pathways for hallucination. Systematic evaluation requires isolating both: which visual conditions trigger perceptual errors, and which integration demands are most vulnerable to failure.

Despite the importance of understanding object hallucination in multi-image contexts, existing evaluation frameworks do not adequately address these two dimensions. Most object hallucination benchmarks [18, 28, 43, 47, 51] are designed for single-image scenarios with binary questions, focusing narrowly on existence and counting. This design cannot reveal how visual difficulty factors (perceptual challenges, contextual biases) systematically induce hallucination, nor can it assess multi-image reasoning patterns or compositional capabilities (attributes, spatial relations) that require fine-grained object understanding. While general multi-image benchmarks [11, 16, 26, 33, 37, 50] evaluate overall reasoning, they lack the controlled manipulation of visual factors needed to diagnose object hallucination vulnerabilities. Recent work such as MIHBench [27] explores multi-image hallucination but examines only image quantity as an adversarial factor in ablation studies, without systematically varying visual difficulty or distinguishing between reasoning patterns. Consequently, existing benchmarks cannot pinpoint which specific visual conditions or reasoning demands trigger hallucination, nor how

*Equal contribution.
†Corresponding author.

these factors interact.

To narrow these gaps, we introduce the Fine-Grained Multi-Image Object Hallucination (MIOH) benchmark, designed to systematically evaluate both dimensions of **multi-image object hallucination**. MIOH systematically integrates four object-centric tasks (existence, counting, attribute, position) with three reasoning patterns (comprehensive, comparative, selective), enabling compositional evaluation from basic detection to fine-grained property binding. Furthermore, MIOH introduces three controlled adversarial pressures—visual context scale (number of images), perceptual difficulty (small/occluded objects), and contextual bias (misleading co-occurrence priors)—that systematically vary visual complexity while measuring performance across reasoning patterns. This design enables diagnostic analysis by separately measuring performance under different visual conditions (perceptual difficulty, contextual bias, image scale) and across different reasoning patterns (comprehensive, comparative, selective).

Our contributions are summarized as follows:

- We introduce MIOH, the first benchmark to systematically assess **object hallucination in multi-image contexts** with fine-grained diagnostic capabilities across visual complexity and reasoning demands.
- We define **three multi-image reasoning patterns** (comprehensive, comparative, selective) and instantiate them across four foundational object-centric tasks (existence, counting, attribute, position), enabling diagnosis of which reasoning capability fails under hallucination pressures.
- We design **three controllable adversarial pressures** (visual context scale, perceptual difficulty, contextual bias) that systematically exacerbate hallucination, allowing fine-grained analysis of when and why MLLMs fail in multi-image scenarios.

2. Related Work

Multimodal Large Language Models (MLLMs). Following the success of LLMs, MLLMs have rapidly evolved through visual instruction tuning, utilized by LLaVA [30] and extended by InstructBLIP [8] and MiniGPT-4 [62]. Early MLLMs face challenges in cross-image reasoning due to limitations in visual token processing and inter-image semantic modeling [2, 4, 8, 21, 25]. Recent work has enabled multi-image understanding [16, 20, 24, 36, 57]; Mantis [16], LLaVA-NeXT-Interleave [24], and Idefics3 [20] leverage large-scale image-text data at training, and Qwen2.5-VL [3], InternVL3.5 [54], and Gemini-2.5-Pro [6] demonstrate powerful cross-image reasoning capabilities.

Object Hallucination in MLLMs. Object hallucination, defined as MLLMs generating plausible but inaccurate object descriptions inconsistent with visual inputs, remains a critical challenge [7, 43]. Systematic analysis has pinpointed causes across the MLLM pipeline: data-related issues such as statistical biases in training data [28], limitations of vision encoder in fine-grained semantics [48], insufficient modality alignment [31], and inherited LLM biases such as weak context attention [53]. Recent studies reveal additional visual vulnerabilities, *e.g.*, perception of small or occluded objects [58], contextual bias due to object co-occurrence patterns [28] and semantic similarities [23]. MLLMs are also linguistically susceptible to sycophantic alignment with user beliefs and context hijacking from misleading narratives [61]. Mitigation strategies, *e.g.*, data augmentation [44], preference optimization [59], and inference-time interventions [14, 60], have primarily targeted single image scenarios. Multi-image contexts amplify hallucination challenges, requiring models not only to recognize objects accurately, but to track them and maintain contextual consistency across distinct images [56]. This increased complexity requires a new benchmark tailored to assess object hallucination on multiple images.

Benchmarks for MLLMs. Early MLLM benchmarks focus on single-image scenarios across various tasks including visual question answering, reasoning, and compositional understanding [10, 12, 22, 29, 32]. Recently, several benchmarks [11, 16, 26, 33, 37, 50] assess general reasoning capabilities across multiple images, though none of them specifically target object hallucination.

In parallel, object hallucination has been addressed through specialized benchmarks, including discriminative approaches using binary questions [13, 28] and generative approaches directly assessing free-form descriptions [18, 43, 47, 51]. They typically focus on existence and simple counting tasks, with limited coverage of attributes or spatial relations. Also, they predominantly employ simple binary questions or captioning tasks, confined to single-image settings. Recently, MIHBench[27] has also explored hallucination in multi-image settings. However, it focuses mainly on binary existence questions and varies only the number of images, without modeling broader visual difficulty factors or diverse object-centric tasks. Our benchmark is complementary to this direction, offering a more fine-grained diagnostic space across multiple tasks, reasoning patterns, and adversarial conditions.

3. Overall Design of our MIOH Benchmark

We introduce the Multi-Image Object Hallucination (MIOH) benchmark, designed to systematically evaluate object hallucination in MLLMs under multi-image contexts. MIOH is structured around two complementary dimensions: (1) **multi-image reasoning patterns** that probe different integration capabilities (Sec. 3.1), and (2) **adversarial pressures** that systematically challenge perceptual

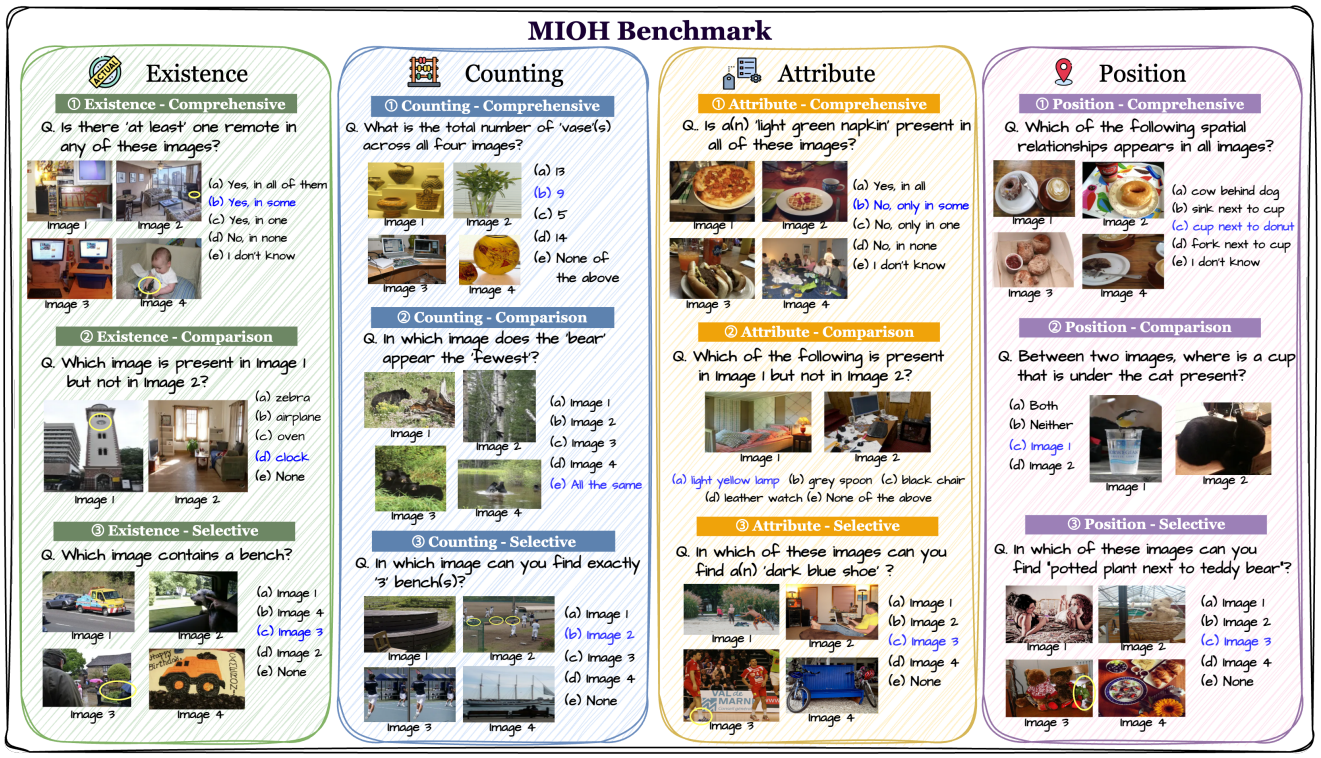


Figure 1. **Overview of our MIOH Benchmark.** MIOH evaluates object hallucination in multi-image contexts across four core tasks: existence, counting, attribute, and position. Each task includes three question types (comprehensive, comparative, and selective) designed to probe different aspects of multi-image reasoning capabilities.

and contextual robustness (Sec. 3.3). By crossing four fundamental object-centric tasks with three reasoning patterns under varying adversarial conditions, MIOH enables fine-grained diagnosis of when and why MLLMs hallucinate.

3.1. Multi-Image Reasoning Patterns

Motivation. Multi-image understanding requires fundamentally different reasoning capabilities compared to single-image settings. While single-image benchmarks test whether models can recognize objects in isolation, multi-image scenarios demand that models *synthesize, compare, and retrieve* information across multiple visual contexts. We identify three core reasoning patterns that capture these distinct capabilities, each placing unique demands on the model's information integration mechanisms.

Comprehensive Reasoning. requires models to aggregate information across all images to form a holistic judgment. Questions in this category ask about collective properties spanning the entire image set, such as "Does any image contain a zebra?" or "What is the total count of cars across all images?" This pattern tests the model's ability to maintain a unified representation of information distributed across multiple scenes—a capability essential for summarization and global understanding tasks. Failures in comprehensive reasoning indicate difficulties in information aggregation or token limitations.

Comparative Reasoning. requires models to identify differences between specific images. Questions such as "Which image contains more cars?" demand precise cross-image attention and the ability to maintain separate representations for different scenes while comparing them. This is critical for change detection, progress monitoring, and contrastive analysis. Failures suggest weaknesses in maintaining distinct object representations across contexts.

Selective Reasoning. requires models to retrieve a specific image that matches given criteria from a set of candidates. Questions like "In which image are exactly three zebras present?" test the ability to perform targeted retrieval while filtering out irrelevant information. This is essential for search and localization tasks in multi-image contexts. Failures indicate poor image-level indexing or inability to isolate relevant visual evidence from distractors.

Diagnostic Value. These three patterns are not merely different question formats—they probe distinct failure modes. Uniform failure across all patterns suggests a fundamental bottleneck in object recognition (perceptual failure). Conversely, failure in *selective* reasoning while succeeding in *comprehensive* reasoning points to a specific integration failure in retrieval and localization. This systematic eval-

uation allows us to diagnose *which* integration capability breaks down under specific adversarial conditions.

3.2. Object-Centric Tasks

We apply these three reasoning patterns to four fundamental object-centric tasks that represent core visual understanding capabilities: **Existence** (verifying object presence/absence), **Counting** (enumerating objects), **Attribute** (binding visual properties to objects, e.g., "red car"), and **Position** (binding spatial relationships, e.g., "dog next to a cat"). While Existence and Counting have been the primary focus of object hallucination benchmarks [5, 18, 28, 35, 43, 52], all four tasks form the foundation for assessing object-centric capabilities of MLLMs [10, 17, 32, 41, 47, 49, 50].

From Tasks to Question Types. By crossing four tasks with three reasoning patterns, we create a comprehensive evaluation space. However, not all combinations are equally meaningful or distinct. For example, in the Counting task, comprehensive reasoning ("total count across all images") naturally leads to multiple variants depending on whether we ask for exact counts, comparisons of counts, or presence of specific quantities. Through careful design, we developed **26 distinct question types** that systematically cover the task and reasoning space while ensuring each type probes a unique aspect of multi-image understanding. Tab. 1 presents the complete taxonomy with representative templates for each type. Complete examples are in Fig. 1 and Sec. C.

Table 1. **MIOH Question Type Taxonomy.** We design 26 question types by crossing 4 object-centric tasks with 3 reasoning patterns. Each type is illustrated with a template showing its structure.

Type	Template
	Existence
Comprehensive	Is there at least one {object} in any of these images?
Comprehensive	Is a {object} present in all of these images?
Comprehensive	Which object appears in at least one image?
Comprehensive	Which object appears in all images?
Selective	In which image does a {object} appear?
Comparative	Which of the two images contains a {object}?
Comparative	Which object is in Image 1 but not Image 2?
	Attribute
Comprehensive	Is a "{attribute} {object}" present in any image?
Comprehensive	Is a "{attribute} {object}" present in all images?
Comprehensive	Which attribute–object pair appears in one image?
Comprehensive	Which attribute–object pair appears in all images?
Selective	In which image is "{attribute} {object}" found?
Comparative	Which of the two images "{attribute} {object}" present?
Comparative	Which attribute–object pair is in Image 1 but not Image 2?
	Counting
Comprehensive	What is the total number of "{object}" across images?
Comprehensive	Which object appears {count} times across images?
Comprehensive	In how many images is a "{object}" present?
Selective	In which image are {count} "{object}" found?
Comparative	In which image does the "{object}" appear most/least?
	Position
Comprehensive	Is a {subject} {relation} a {anchor} present in any image?
Comprehensive	Is a {subject} {relation} a {anchor} present in all images?
Comprehensive	Which object with spatial relationship appears in one image?
Comprehensive	Which object with spatial relationship appears in all images?
Selective	In which image is a {subject} {relation} a {anchor}?
Selective	Where is "{subject} {relation} {anchor}" present?
Comparative	Which object with spatial relationship is in Image 1 but not Image 2?

3.3. Adversarial Pressures

To systematically probe model vulnerabilities, we introduce three adversarial factors that create challenging but realistic evaluation conditions. These factors target distinct failure mechanisms: **visual context scale** tests integration capacity as the number of images increases, **perceptual difficulty** challenges feature extraction from small or occluded objects, and **contextual bias** probes susceptibility to misleading co-occurrence priors.

Visual Context Scale (Number of Images). Inspired by findings that MLLMs struggle to identify information across large image sets (the "Visual Haystack" problem [56]), we systematically vary the **Number of input Images (NI)** among $\{2, 4, 8, 10\}$ for the same underlying question. This tests the model's *integration capacity*—as the visual context expands with more images, models must maintain accurate object representations across an increasing number of visual scenes.

Perceptual Difficulty (Hard Positive). Small or partially occluded objects are inherently harder to detect [34, 55, 58]. We curate **Hard Positive (HP)** examples using two complementary approaches: (a) *Rule-based filtering* selects images where target objects have small bounding boxes or high occlusion ratios; (b) *CLIP-based semantic filtering* identifies cases where CLIP similarity between the image and text prompt "A photo of [object]" is abnormally low, indicating perceptual ambiguity. These examples test whether models can extract accurate features under perceptually challenging conditions—a necessary prerequisite for any downstream reasoning.

Contextual Bias (Hard Negative). Strong contextual priors can mislead models into hallucinating objects that are contextually plausible but visually absent [9, 23, 28]. For example, a kitchen scene may activate strong priors for "frying pan," leading to false positives. We construct **Hard Negative (HN)** examples by: (a) estimating co-occurrence probabilities $P(\text{object}_A|\text{object}_B)$ from COCO training data and selecting images containing high-probability co-occurring objects but missing the target; (b) applying CLIP-based semantic confusion to find images with high visual-text similarity despite object absence. These examples test whether models rely on contextual shortcuts rather than grounded visual evidence.

3.4. Implementation details

Dataset Selection. To ensure annotation quality, we use three complementary datasets: **COCO-ReM** [46] (existence, counting) addresses COCO's incomplete masks and missing instances via systematic re-annotation; **PACO** [42] (attributes) provides standardized attribute labels across ob-

Figure 2. **Representative examples from MIOH.** (Top) Existence-Comprehensive questions across Easy, Hard Negative, and Hard Positive scenarios. (Bottom) Counting-Selective questions with varying difficulty levels. Additional examples are provided in Sec. C.

ject categories; **SVG** [39] (spatial relationships) offers complete scene-level annotations, unlike Visual Genome [19] which averages only 1.5 relationships per subject.

Benchmark Statistics. Three independent annotators validated all questions. The final benchmark comprises 3,484 questions across 11,732 images, balanced across tasks, reasoning patterns, and difficulty levels. Full construction details, filtering criteria, and statistics are in Sec. A.

4. Benchmark Results and Discussion

We conduct a comprehensive comparative study using our MIOH over the state-of-the-art MLLMs, including GPT-5 [45] and Gemini-2.5-Pro [6]. Among open-source models, we choose the LLaVA [24] series, the Qwen [3] series, InternVL [54], Phi-4-multimodal [1], MiniCPM-V [57], Ovis-2.5 [36], and Mantis-8B [16]. For reproducibility, the decoding temperature is set to 0 for all experiments. All experiments were conducted on four NVIDIA A6000 GPUs.

4.1. Overall Performance

We first present overall performance results demonstrating the extent of object hallucination challenges across different model categories and tasks. Tab. 2 reports the comprehensive evaluation results over 29 models, revealing that multi-image object hallucination remains as significant challenge, with an overall average accuracy of only **36.1%**. Not surprisingly, a clear performance gap is measured between proprietary and open-source models. The leading models, Gemini-2.5 Pro (64.4%) and GPT-5 (63.1%), set the state-of-the-art, but are still far from perfect. Top-performing open-source models, such as Qwen2-VL-7B (49.1%) and MiniCPM-V-2.6 (48.5%), demonstrate strong capabilities but lag behind the frontier models.

Beyond the overall gap, performance is shaped by three key factors—reasoning pattern, task type, and adversarial pressure—each revealing distinct failure modes, which we analyze in the following subsections.

4.2. Reasoning Patterns

Our analysis reveals that MLLM performance is governed not only by the task but also, fundamentally, by the required reasoning pattern. As shown in Fig. 3, performance varies significantly across Comprehensive, Comparison, and Selection patterns, demonstrating that existing object hallucination benchmarks, constrained to a few question styles, cannot properly assess object hallucination across diverse multi-image settings.

Overall, Comprehensive emerges as the most manageable reasoning pattern, with basic Existence questions yielding the highest performance. Comparing this densely high-performing row against the rest of the heatmap reveals that models are heavily biased towards simple, presence-verifying queries. Evaluating models solely within this narrow scope inevitably overestimates their true robustness.

On average, Selection(34.2%) emerges as the most challenging reasoning pattern, underperforming both Comprehensive and Comparison. This difficulty becomes especially pronounced when paired with compositional tasks such as Attribute(26.7%), indicating that models often recognize that an object-attribute pair exists within the image set but fail to identify which specific image contains it—a distinct form of grounding failure in multi-image settings.

Model	Existence					Counting				
	Easy	HN	HP	NI	**Avg**	Easy	HN	HP	NI	**Avg**
Overall	62.4	51.9	51.5	30.0	**49.4**	24.3	29.5	27.0	20.5	**25.4**
GPT-5	91.4	88.6	78.4	55.3	**78.4**	55.0	54.3	51.5	35.7	**49.1**
Gemini-2.5 Pro	83.0	81.5	80.7	56.4	**75.4**	64.4	66.4	59.4	40.0	**57.5**
Qwen2-VL-2B	58.3	53.5	45.9	32.7	**47.6**	21.7	23.3	18.6	21.4	**21.2**
Qwen2.5-VL-3B	80.7	62.5	55.5	44.7	**60.8**	23.3	26.7	19.6	17.1	**21.7**
Qwen2-VL-7B	87.3	75.5	76.4	30.7	**67.5**	28.9	36.2	21.6	17.1	**26.0**
Qwen2.5-VL-7B	84.0	76.0	73.6	28.0	**65.4**	33.3	39.7	17.5	21.4	**28.0**
LLaVA-v1.6 (Mistral-7B)	33.0	47.0	36.8	-	**38.9**	20.6	27.6	18.6	-	**22.2**
LLaVA-Interleave (Qwen-0.5B)	36.3	31.5	28.6	18.7	**28.8**	22.8	28.4	24.7	14.3	**22.6**
LLaVA-Interleave (Qwen-7B)	46.7	44.0	65.5	30.7	**46.7**	20.6	25.0	28.9	24.3	**24.7**
LLaVA-Interleave (Qwen-7B-DPO)	67.7	44.0	55.9	31.3	**49.7**	21.1	25.9	30.9	27.1	**26.3**
LLaVA-OneVision (Qwen2-0.5B-SI)	27.7	11.5	37.3	40.0	**29.1**	13.9	21.6	18.6	12.9	**16.7**
LLaVA-OneVision (Qwen2-0.5B-OV)	57.3	23.5	33.6	22.0	**34.1**	16.7	26.7	16.5	18.6	**19.6**
LLaVA-OneVision (Qwen2-7B-SI)	68.3	46.5	40.9	36.0	**47.9**	17.2	24.1	25.8	25.7	**23.2**
LLaVA-OneVision (Qwen2-7B-OV)	83.3	78.0	57.7	24.7	**60.9**	18.3	28.4	29.9	18.6	**23.8**
LLaVA-OneVision (Qwen2-7B-OV-Chat)	82.7	77.0	43.6	26.7	**57.5**	20.0	30.2	27.8	21.4	**24.9**
InternVL3.5-1B	44.0	41.0	30.5	42.0	**39.4**	21.7	21.6	19.6	21.4	**21.1**
InternVL3.5-2B	51.7	37.5	36.8	26.0	**38.0**	15.0	28.4	16.5	17.1	**19.3**
InternVL3.5-4B	59.0	71.0	60.0	22.7	**53.2**	18.9	27.6	34.0	17.1	**24.4**
InternVL3.5-8B Pretrained	72.0	62.0	59.5	30.0	**55.9**	22.8	26.7	27.8	14.3	**22.9**
InternVL3.5-8B Instruct	76.0	59.5	75.5	30.7	**60.4**	25.6	29.3	29.9	8.6	**23.3**
InternVL3.5-8B MPO	76.7	60.0	76.4	18.0	**57.8**	26.1	29.3	28.9	10.0	**23.6**
InternVL3.5-8B	34.0	55.5	34.1	20.0	**35.9**	30.0	37.1	32.0	11.4	**27.6**
Mantis-8B (CLIP-Llama3)	45.7	57.0	36.8	18.7	**39.5**	22.8	25.9	21.6	30.0	**25.1**
Mantis-8B (SIGLIP-Llama3)	47.0	45.0	42.7	24.7	**39.8**	20.6	25.0	33.0	22.9	**25.4**
MiniCPM-Llama3-V-2.5	35.0	16.5	25.5	4.7	**20.4**	4.4	5.2	8.2	5.7	**5.9**
MiniCPM-V-2.6	86.3	75.5	73.6	38.7	**68.5**	21.1	35.3	38.1	18.6	**28.3**
Ovis2.5-2B	72.0	23.5	47.3	34.0	**44.2**	23.9	25.9	18.6	25.7	**23.5**
Ovis2.5-9B	78.0	23.0	51.8	28.7	**45.4**	28.9	27.6	48.5	25.7	**32.7**
Phi-4-multimodal	45.3	38.0	33.6	23.3	**35.1**	25.0	25.0	17.5	28.6	**24.0**

Model	Attribute					Position					Overall
	Easy	HN	HP	NI	**Avg**	Easy	HN	HP	NI	**Avg**	**Avg**
Overall	37.5	32.9	31.3	26.9	**32.4**	46.9	35.2	45.2	22.2	**37.3**	**36.1**
GPT-5	65.8	62.4	57.0	45.8	**57.8**	88.5	70.3	75.5	34.1	**67.1**	**63.1**
Gemini-2.5 Pro	62.9	57.1	64.3	47.5	**57.9**	80.8	66.4	81.4	37.6	**66.6**	**64.4**
Qwen2-VL-2B	42.3	34.2	50.0	25.0	**37.9**	46.2	38.4	74.4	22.0	**45.3**	**38.0**
Qwen2.5-VL-3B	47.3	38.8	47.1	26.7	**40.0**	73.8	50.4	72.1	22.0	**54.6**	**44.3**
Qwen2-VL-7B	58.8	36.2	44.3	28.3	**41.9**	86.2	56.8	79.5	21.7	**61.1**	**49.1**
Qwen2.5-VL-7B	52.7	42.1	38.1	26.7	**39.9**	76.2	52.4	60.5	21.4	**52.6**	**46.5**
LLaVA-v1.6 (Mistral-7B)	30.4	21.7	24.3	-	**25.4**	29.2	25.2	31.2	-	**28.5**	**28.8**
LLaVA-Interleave (Qwen-0.5B)	41.5	31.7	31.4	17.5	**30.5**	15.0	17.6	40.9	34.5	**27.0**	**27.2**
LLaVA-Interleave (Qwen-7B)	27.3	27.5	29.0	15.8	**24.9**	31.2	24.8	32.1	21.4	**27.4**	**30.9**
LLaVA-Interleave (Qwen-7B-DPO)	28.5	28.8	28.6	13.3	**24.8**	37.1	26.4	30.2	18.5	**28.1**	**32.2**
LLaVA-OneVision (Qwen2-0.5B-SI)	22.3	17.9	27.1	12.5	**20.0**	18.3	15.6	34.9	16.8	**21.4**	**21.8**
LLaVA-OneVision (Qwen2-0.5B-OV)	23.1	17.1	28.1	22.5	**22.7**	19.2	21.2	33.5	22.9	**24.2**	**25.2**
LLaVA-OneVision (Qwen2-7B-SI)	24.2	21.2	24.3	28.3	**24.5**	37.5	30.4	42.8	27.1	**34.4**	**32.5**
LLaVA-OneVision (Qwen2-7B-OV)	30.4	24.6	27.1	23.3	**26.4**	47.9	24.8	43.7	21.1	**34.4**	**36.4**
LLaVA-OneVision (Qwen2-7B-OV-Chat)	30.8	25.0	26.7	22.5	**26.2**	48.3	25.6	43.3	21.0	**34.5**	**35.8**
InternVL3.5-1B	28.8	24.2	27.1	17.5	**24.4**	57.1	37.2	25.1	32.7	**38.0**	**30.7**
InternVL3.5-2B	40.4	37.9	15.2	35.0	**32.1**	60.0	43.6	13.5	18.4	**33.9**	**30.8**
InternVL3.5-4B	42.3	39.2	34.8	35.8	**38.0**	31.7	18.8	57.2	19.2	**31.7**	**36.8**
InternVL3.5-8B Pretrained	40.8	37.5	19.5	38.3	**34.0**	37.5	32.0	37.2	28.1	**33.7**	**36.6**
InternVL3.5-8B Instruct	38.1	34.6	18.1	39.2	**32.5**	35.4	33.2	37.7	19.8	**31.5**	**36.9**
InternVL3.5-8B MPO	37.7	42.9	14.8	39.2	**33.6**	36.2	34.0	20.0	18.5	**27.2**	**35.5**
InternVL3.5-8B	34.6	27.5	28.6	37.5	**32.0**	55.8	36.4	35.8	35.0	**40.8**	**34.1**
Mantis-8B (CLIP-Llama3)	24.6	24.6	24.8	20.0	**23.5**	25.8	25.2	31.6	20.6	**25.8**	**28.5**
Mantis-8B (SIGLIP-Llama3)	26.2	29.2	26.2	20.0	**25.4**	31.2	22.8	28.8	19.5	**25.6**	**29.0**
MiniCPM-Llama3-V-2.5	31.2	30.8	28.6	9.2	**24.9**	44.6	44.8	53.5	15.0	**39.5**	**22.7**
MiniCPM-V-2.6	53.8	43.3	33.8	21.7	**38.2**	79.6	62.0	69.3	25.2	**59.0**	**48.5**
Ovis2.5-2B	26.2	42.9	34.3	21.7	**31.3**	35.0	20.8	40.0	31.0	**31.7**	**32.7**
Ovis2.5-9B	39.6	27.9	33.8	26.7	**32.0**	51.2	24.4	44.2	37.6	**39.4**	**37.3**
Phi-4-multimodal	34.6	26.7	21.0	35.0	**29.3**	42.5	38.4	39.5	22.7	**35.8**	**31.0**

Table 2. **MLLM performance on MIOH benchmark.** HN: Hard Negatives, HP: Hard Positives, NI: Number of Images.

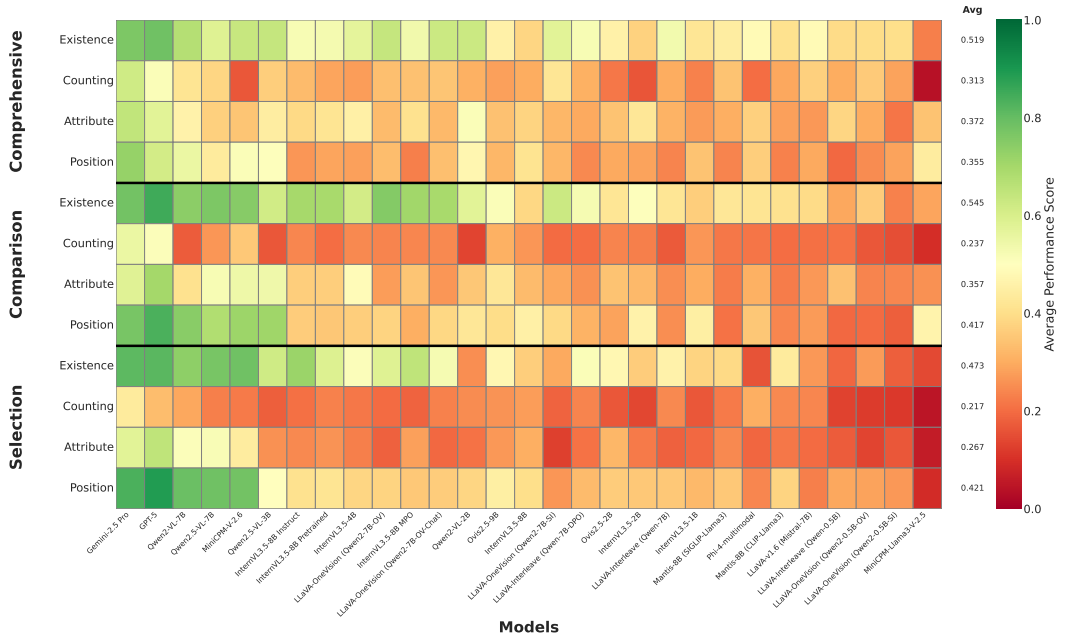


Figure 3. Performance heatmap across multi-image reasoning patterns (Comprehensive, Comparison, Selection) and object-centric tasks (Existence, Counting, Attribute, Position).

4.3. Task-Specific Vulnerability

We conduct detailed analyses to identify specific failure patterns and characterize how different adversarial pressures affect model behavior in multi-image scenarios. As reported in Tab. 2, the overall task-level accuracy reveals a clear hierarchy of difficulty. Existence (49.4%) emerges as the most manageable task for most models, suggesting a basic level of object recognition. In stark contrast, counting (25.4%) is a critical failure point across the board, highlighting a fundamental weakness in quantitative reasoning under multi-image contexts. Attribute (32.4%) and position (37.3%) tasks reveal significant difficulty as well, underscoring the challenges of compositional understanding.

Existence. While Existence is the highest-scoring task, model performance is brittle. The accuracy on easy samples (62.4%) consistently drops on perceptually challenging objects (HP, 51.5%) or on contextually misleading scenes (HN, 51.9%). This indicates that models' understanding of object presence is heavily reliant on clear, unambiguous visual cues and can be easily disrupted, leading to false negative hallucinations.

Counting. The average accuracy of 25.4% confirms that counting is a core deficiency. Our sub-task analysis reveals this is not limited to one failure mode; models struggle with both aggregating counts across images and locating an image with a specific number of objects. This points to a fundamental inability in quantitative grounding, especially on multiple scenes.

Attribute and Position. These tasks require binding objects to their properties or spatial relations, a challenge of compositional reasoning. The low accuracy in the attribute task (32.4%) suggests that MLLMs may recognize an object and an attribute separately but fail to confirm their visual co-occurrence (*e.g.*, seeing a 'car' and 'red' but hallucinating a 'red car'). Similarly, the position task scores slightly higher (37.3%), but degrades under Hard Negative pressure, indicating that models often tend to overlook complex spatial relationships beyond mere existence.

4.4. Impact of Adversarial Pressures

As reported in Tab. 2, introducing adversarial constraints consistently degrades MLLM performance across all tasks relative to their easy baselines.

Perceptual and Contextual Pressures (HP & HN). Performance on Existence, Attribute, and Position consistently drops under both HP and HN conditions. For instance, Attribute accuracy falls from 37.5% to 31.3% (HP), and Position from 46.9% to 35.2% (HN), confirming model vulnerabilities to misleading visual and contextual cues. In contrast, Counting exhibits only a marginal decline. However, this reflects a *floor effect* rather than genuine robustness; with baseline counting accuracy already near random chance (25.4%), added complexities leave little room for further degradation.

Cognitive Load from Multiple Images (NI). While HP and HN induce moderate decays, increasing the number of input images to 8 (NI) triggers a far more severe collapse. Under this heightened cognitive load, even the fundamental Existence task plummets to 30.0%—a relative drop of exceeding 50% from the easy baseline. Although top-tier models like GPT-5 and Gemini 2.5 Pro manage this ex-

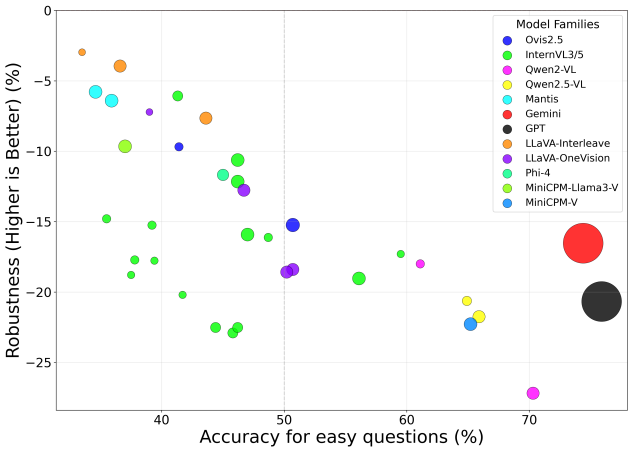


Figure 4. **Accuracy on easy questions and robustness under adversarial conditions in VLMs.** Circle size indicates model size, where Gemini and GPT size is based on estimation.

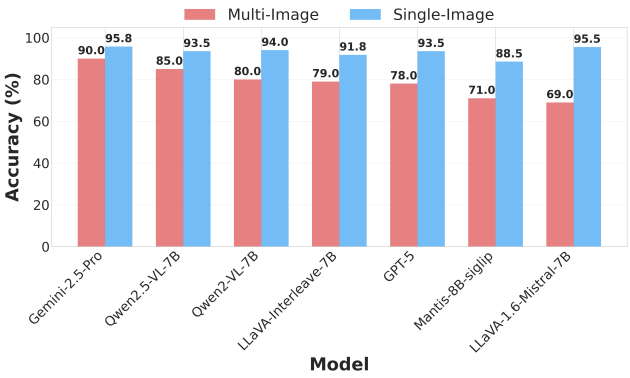


Figure 5. **Multi-image processing consistently degrades object hallucination across all models.** Comparison of accuracy between multi-image comprehensive (red bars) and single-image decomposed (blue bars) evaluation on Existence task.

panded context better than average, their substantial performance drops confirm that scaling multi-image reasoning remains a critical open challenge.

4.5. Accuracy-Robustness Trade-off

As shown in Fig. 4, our analysis uncovers a fundamental trade-off between achieving high accuracy on straightforward tasks and maintaining robust performance under adversarial pressures. In other words, a strong baseline performance does not guarantee robustness against object hallucination. Open-source models with higher baseline accuracy often exhibit greater vulnerability to adversarial conditions. The correlation analysis reveals a moderate positive relationship between model size and performance on easy questions, but virtually no correlation between size and robustness, suggesting that simply scaling model parameters does not inherently improve resilience to object hallucination. The top-performing open-source models on easy questions—Qwen2-VL-7B and Qwen2.5-VL-7B—experience substantial robustness drops of -27.2% and -21.8%, respectively, indicating that high capability models might be more susceptible to the adversarial pressures we designed. This finding suggests that the ability to handle straightforward multi-image tasks does not guarantee robustness against object hallucination.

4.6. Multi-Image Context as a Hallucination Amplifier: An Ablation Study

To isolate the impact of multi-image processing on object hallucination, we conduct a controlled ablation study focusing on the Existence task. Specifically, we compare two evaluation approaches for identical visual content: **comprehensive** questions that require synthesizing information across all images simultaneously ("Is there an OBJECT in any of these IMAGEs?") *vs.* **decomposed** questions that mirror the traditional single-image setting by asking each image separately ("Is there a OBJECT in IMAGE?") and combining the answers to determine overall presence. This design isolates whether multi-image contexts introduce systematic errors beyond simple accumulation of individual image processing mistakes.

The results in Fig. 5 reveal that single-image processing substantially outperforms simultaneous multi-image analysis, consistently across all models and scales. This indicates that multi-image contexts significantly amplify object hallucination beyond what error accumulation alone would predict, pointing to cross-image integration as a primary source of failure. The consistency of this penalty suggests that current MLLM training paradigms fail to adequately address object hallucination in multi-image reasoning.

5. Conclusion

We introduce MIOH, a benchmark that systematically evaluates object hallucination in multi-image contexts across four object-centric tasks (Existence, Counting, Attribute, Position) with three reasoning patterns (Comprehensive, Comparative, Selective) under three adversarial pressures (Visual Context Scale, Perceptual Difficulty, Contextual Bias). Through evaluation of 29 models, we demonstrate that current MLLMs—including GPT-5 and Gemini-2.5-Pro—exhibit fundamental limitations in multi-image integration. Our fine-grained diagnostic framework reveals that hallucination stems not merely from perceptual failures but from integration-stage breakdowns when maintaining object representations across multiple images. MIOH provides a foundation for targeted improvements in multi-image understanding and serves as a critical evaluation tool for developing more reliable multimodal AI systems.

Acknowledgments

This work was also supported by the SOFT Foundry Institute at SNU, Samsung Electronics, Youlchon Foundation, National Research Foundation of Korea (NRF) grants (RS-2021-NR05515, RS-2024-00336576, RS-2023-0022663, RS-2025-25399604, RS-2024-00333484), and the Institute for Information & Communication Technology Planning & Evaluation (IITP) grants (RS-2022-II220264, RS-2024-00353131) funded by the Korean government.

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:2503.01743*, 2025. 5

[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. In *NeurIPS*, 2022. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv:2502.13923*, 2025. 2, 5

[4] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *CVPR*, 2024. 2

[5] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *NeurIPS*, 2024. 4

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*, 2025. 2, 5

[7] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv:2210.07688*, 2022. 2

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 2

[9] Shounak Datta and Dhanasekar Sundararaman. Evaluating hallucination in large vision-language models based on context-aware object similarities. *arXiv:2501.15046*, 2025. 4

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023. 2, 4

[11] 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 *ECCV*, 2024. 1, 2

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2

[13] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024. 2

[14] Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects? In *NAACL*, 2025. 2

[15] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. i

[16] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: Interleaved multi-image instruction tuning. *arXiv:2405.01483*, 2024. 1, 2, 5

[17] Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, 2024. 4

[18] Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *CVPR*, 2024. 1, 2, 4

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5, i

[20] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv:2408.12637*, 2024. 2

[21] Hyun Lee, Hyemin Jeong, Yejin Kim, Hyungwook Choi, Hyunsoo Cho, Sookyung Kim, and Joonseok Lee. A more word-like image tokenization for mllms. In *CVPR*, 2026. 2

[22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*, 2023. 2

[23] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *NeurIPS*, 2024. 2, 4

[24] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-NeXT-interleave: Tackling multi-image, video, and 3D in large multimodal models. *arXiv:2407.07895*, 2024. 2, 5

[25] 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 *ICML*, 2023. 2

[26] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal LLMs to follow zero-shot demonstrative instructions. *arXiv:2308.04152*, 2023. 1, 2

[27] Jiale Li, Mingrui Wu, Zixiang Jin, Hao Chen, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. MIHBench: Benchmarking and mitigating multi-image hallucinations in multimodal large language models. In *ACM MM*, 2025. 1, 2

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*, 2023. 1, 2, 4

[29] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 2

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024. 2, 4

[33] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *Advances in Neural Information Processing Systems*, 37: 8698–8733, 2024. 1, 2

[34] Zhaochen Liu, Kaiwen Gao, Shuyi Liang, Bin Xiao, Limeng Qiao, Lin Ma, and Tingting Jiang. Beyond the visible: Benchmarking occlusion perception in multimodal large language models. *arXiv:2508.04059*, 2025. 4

[35] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv:2310.05338*, 2023. 4

[36] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2.5 technical report. *arXiv:2508.11737*, 2025. 2, 5

[37] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv:2408.02718*, 2024. 1, 2

[38] Yannic Neuhaus and Matthias Hein. RePOPE: Impact of annotation errors on the POPE benchmark. *arXiv:2504.15707*, 2025. i

[39] Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, et al. Synthetic visual genome. In *CVPR*, 2025. 5, i

[40] Genevieve Patterson and James Hays. COCO attributes: Attributes for people, animals, and objects. In *ECCV*, 2016. i

[41] Han Qiu, Jiaxing Huang, Peng Gao, Qin Qi, Xiaoqin Zhang, Ling Shao, and Shijian Lu. LongHalQA: Long-context hallucination evaluation for multimodal large language models. *arXiv:2410.09962*, 2024. 4

[42] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. PACO: Parts and attributes of common objects. In *CVPR*, 2023. 4, i

[43] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv:1809.02156*, 2018. 1, 2, 4

[44] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *arXiv:2405.18654*, 2024. 2

[45] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 5

[46] Shweta Singh, Aayan Yadav, Jitesh Jain, Humphrey Shi, Justin Johnson, and Karan Desai. Benchmarking object detectors with COCO: A new path forward. In *ECCV*, 2024. 4, i

[47] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv:2309.14525*, 2023. 1, 2, 4

[48] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 2

[49] Andrés Villa, Juan Léon, Alvaro Soto, and Bernard Ghanem. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. In *CVPR*, 2025. 4

[50] 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. *arXiv:2406.09411*, 2024. 1, 2, 4

[51] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. AMBER: An LLM-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv:2311.07397*, 2023. 1, 2

[52] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, 2024. 4

[53] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. CogVLM: Visual expert for pretrained language models. In *NeurIPS*, 2024. 2

[54] 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:2508.18265*, 2025. 2, 5

[55] Wenbo Wei, Jun Wang, and Abhir Bhalerao. Coco-olac: A benchmark for occluded panoptic segmentation and image understanding. In *ICASSP*, 2025. 4

[56] Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv:2407.13766*, 2024. 2, 4

[57] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level mllm on your phone. *arXiv:2408.01800*, 2024. 2, 5

[58] Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. Exploring perceptual limitation of multimodal large language models. *arXiv:2402.07384*, 2024. 2, 4

[59] Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. *NeurIPS*, 2024. 2

[60] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *ICML*, 2025. 2

[61] Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, Qi Guo, and Yunji Chen. Towards analyzing and mitigating sycophancy in large vision-language models. *arXiv:2408.11261*, 2024. 2

[62] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 2

Fine-Grained Multi-Image Object Hallucination Benchmark

Supplementary Material

A. Detailed Benchmark Construction

In this section, we provide comprehensive details on our benchmark construction process, including: the curation and selection criteria for our datasets (Sec. A.1), the metadata construction methodology (Sec. A.2), the automated question generation framework (Sec. A.3), and the adversarial benchmark design strategy (Sec. A.4).

A.1. Dataset Curation

A.1.1. COCO-ReM

Annotation Quality Issues in Original COCO. The original COCO dataset annotations contain significant gaps that make them unsuitable for reliable object hallucination evaluation. These issues include incomplete object masks, missing instances, and inaccurate bounding boxes that would introduce systematic errors in our rule-based question generation framework.

COCO-ReM Improvements. COCO-ReM (Refined Masks) [46] addresses these limitations through a comprehensive re-annotation process: (1) *Mask boundary refinement* using the Segment Anything Model (SAM) to improve precision, (2) *Missing instance detection* using advanced detection models to identify previously unlabeled objects, (3) *Label correction* through systematic review and human validation, and (4) *Enhanced object masks and bounding boxes* providing more complete scene coverage.

Impact on Benchmark Quality. As demonstrated in RePOPE [38], high-reliability annotations significantly impact ground truth accuracy, making this a crucial consideration for benchmark design. The enhanced annotation quality in COCO-ReM ensures our existence and counting questions have reliable ground truth labels, substantially reducing false negatives that could arise from missed objects in original COCO annotations.

Object Count Limitations. During validation, we observed that even COCO-ReM's accuracy degrades when object counts exceed certain thresholds. Specifically, images containing more than 10 objects showed decreased annotation reliability. To maintain benchmark integrity, we implemented a conservative approach by limiting counting questions to images with 5 or fewer objects, ensuring high reliability through validated annotations while preserving sufficient complexity for meaningful MLLM evaluation.

A.1.2. PACO

Limitations of Existing Attribute Datasets. While various datasets address object attributes, they suffer from systematic limitations: (1) Original COCO annotations lack standardized attribute labeling across object categories, (2) COCO Attributes [40] provides standardized annotation but suffers from limited diversity in both object categories and attribute types, and (3) Insufficient coverage for comprehensive benchmark construction requiring comparison across diverse objects and attributes.

PACO's Comprehensive Approach. PACO (Parts and Attributes of Common Objects) [42] provides a superior solution through: (1) Broader category coverage spanning a more diverse range of object types, (2) Systematic attribute annotation ensuring consistency across identical objects, (3) Detailed annotation process that identifies constituent object parts and labels their diverse attributes, and (4) Large-scale structured dataset resulting in comprehensive fine-grained object understanding capabilities.

Advantages for Question Generation. PACO's structured approach offers several key benefits: systematic attribute labeling with sufficient scale and diversity to support robust question generation, extensive object-attribute combinations enabling comprehensive evaluation across diverse visual scenarios, standardized annotation framework ensuring consistent evaluation criteria across different object categories, and high-quality ground truth reducing ambiguity in attribute-based question validation.

A.1.3. SVG

Limitations of Existing Spatial Relation Datasets. Existing datasets for spatial relationship evaluation suffer from critical annotation gaps: Visual Genome [19] and GQA [15] provide relation data but have incomplete spatial relationship coverage, missing relationships in ground truth annotations that exist visually but are not labeled, and annotation inconsistencies that reduce reliability for systematic evaluation.

SVG's Multifaceted Approach. SVG (Synthetic Visual Genome) [39] addresses these limitations through comprehensive methodology: object detection integration for accurate entity identification, scene graph enhancement to capture missing relationships, region descriptions providing contextual relationship validation, depth information enabling more accurate spatial reasoning, region masks for precise relationship localization, VQA-based verification for non-spatial relationships to ensure annotation quality, and systematic filtering to reduce incorrect relationship annotations.

Key Advantages for Spatial Evaluation. SVG provides several critical improvements: (1) Richer spatial relation coverage per subject compared to existing datasets, enabling more comprehensive spatial reasoning evaluation,

(2) Comprehensive filtering that systematically reduces incorrect relationships, improving ground truth reliability, (3) Region mask-based verification enabling more reliable relationship identification through visual evidence, and (4) High relation density minimizing the critical impact of missing positional relationships on question accuracy. These enhancements make SVG particularly well-suited for generating position-based questions that can reliably assess MLLM spatial reasoning capabilities in multi-image contexts, where accurate relationship identification becomes even more challenging due to increased visual complexity.

A.2. Metadata Construction

Having established our data sources, we now detail the metadata construction process that enables efficient question generation.

Hierarchical Organization Structure. Our metadata follows a systematic three-level organization: (1) *Task-specific property categorization* where objects are categorized by relevant attributes, relations, or counts, (2) *Difficulty level classification* with Easy/Hard Negative/Hard Positive assignments based on visual and semantic complexity, and (3) *Image identifier mapping* where specific image IDs are linked to categorized objects for efficient retrieval.

Rule-Based Filtering Criteria. We implement several filtering mechanisms: minimum bounding box size requirements to ensure object visibility, occlusion level thresholds based on mask overlap calculations, and image resolution considerations for consistent object detectability across different image qualities. For difficulty classification, we define easy positives/negatives as clear, unambiguous cases with high visibility and minimal contextual confusion, hard positives as present objects with small size, high occlusion, or minimal contextual cues, and hard negatives as absent objects in contexts with high co-occurrence bias or semantic similarity.

CLIP-Based Semantic Similarity Implementation. Our similarity score calculation involves text prompt generation using standardized formats ("A photo of [object]","[attribute] [object]"), image encoding through CLIP visual encoder, cosine similarity computation between text and image embeddings, and threshold determination through empirical validation on representative samples. This metadata system enables rapid question synthesis while maintaining quality through automated filtering based on rule-based criteria, semantic validation using CLIP similarity scores, systematic difficulty categorization across different visual reasoning scenarios, and efficient question generation through pre-computed metadata lookup.

A.3. Question Generation

Using the prepared positive and negative samples per task, we generate multi-image questions that instantiate the three types of questions (comprehensive, comparative, and selective) across multi-image contexts for each of the four core tasks. As all ingredients are ready, this step can be easily automated by rule-based templates; *e.g.*, for Counting, we construct a question by randomly selecting N IMAGEs containing a particular OBJECT, and the sum of COUNT per each example is the right answer. Other tasks follow a similar procedure to generate the question and correct answer. All questions are constructed in multiple-choice (MCQ) format by incorporating incorrect answers. We also include the "None of the above" options to more precisely assess the model's understanding.

A.4. Adversarial Example Construction

Building upon the adversarial pressures described in Sec. 3.3, we detail the implementation procedures for constructing challenging evaluation examples.

A.4.1. Hard Positive Example

We employ two complementary filtering approaches to identify perceptually difficult positive examples: **Rule-based filtering** selects (IMAGE, OBJECT) pairs where the target object exhibits challenging visual characteristics. We filter based on bounding box dimensions relative to image size, segmentation mask area indicating occlusion levels, and spatial positioning within the image frame. This approach captures objects that are inherently difficult to perceive due to size or visibility constraints. **CLIP-based semantic filtering** identifies cases where visual-semantic alignment is weak. We compute similarity scores between image embeddings and text prompts formatted as "A photo of OBJECT" using CLIP. Examples with unexpectedly low similarity scores despite object presence indicate perceptual ambiguity—such as unusual viewpoints, partial visibility, or atypical visual contexts that challenge standard recognition patterns.

A.4.2. Hard Negative Example

We construct misleading negative examples through two strategies: **Co-occurrence-based selection** leverages statistical patterns from COCO training data. We compute pairwise co-occurrence probabilities $P(\text{object}_A|\text{object}_B)$ across all object categories. For a target object, we select images containing frequently co-occurring objects while the target itself is absent, creating scenarios where strong contextual priors may mislead models into false positive predictions. **CLIP-based semantic confusion** identifies images that exhibit high visual-semantic similarity with target object prompts despite the object's absence. We compute CLIP similarity between images and "A photo of OBJECT" prompts for absent objects, selecting cases with high similarity scores. These examples represent strong false association triggers where visual context strongly suggests object presence without actual visual evidence.

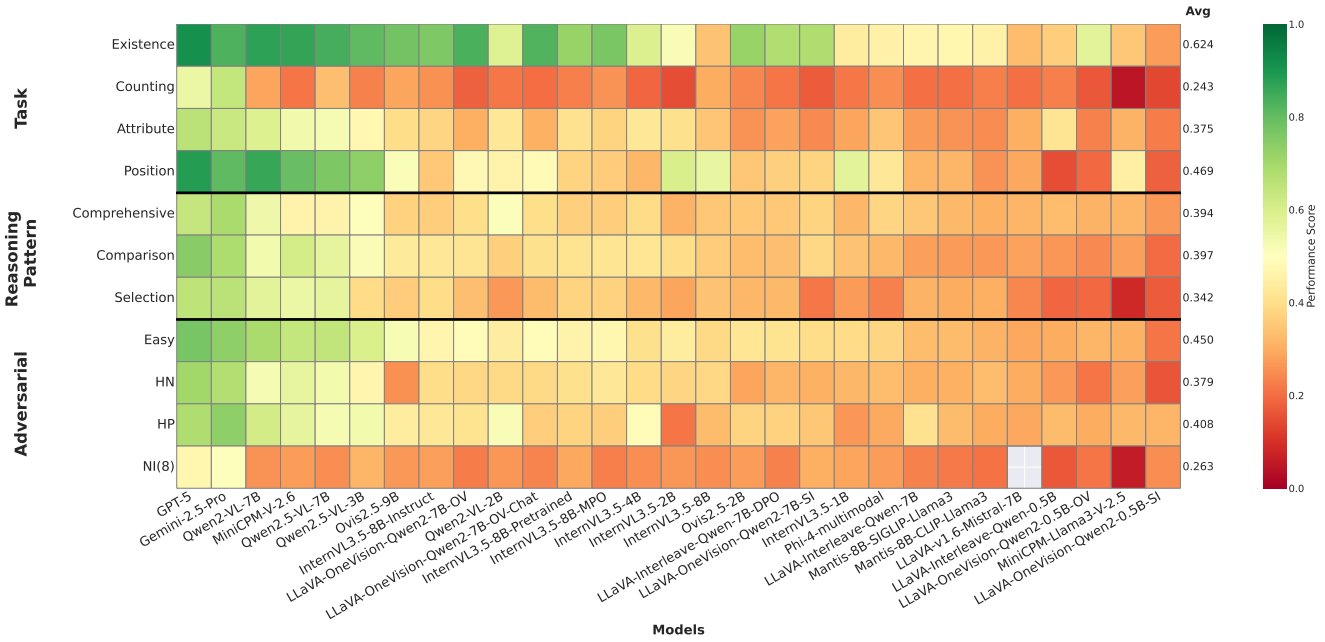


Figure I. **Aggregated Performance Heatmap across Task, Reasoning Pattern, and Adversarial Pressure Dimensions.**

A.4.3. Difficulty Level Integration.

During question generation, we systematically control difficulty by varying the proportion of challenging examples. For a question requiring N images, we can include anywhere from 0 to N hard positive or hard negative examples, with difficulty increasing as the proportion of challenging examples approaches N.

A.4.4. Quality Assurance and Validation

To ensure benchmark reliability, we conducted comprehensive manual validation and systematic sampling to construct a balanced evaluation set. Our automated generation pipeline initially produced over 26,000 questions across all task types and reasoning patterns.

Each question underwent independent review by three annotators with expertise in computer vision and visual reasoning tasks. Annotators assessed: (1) ground truth correctness based on visual evidence, (2) question clarity and lack of ambiguity, (3) validity and distinctiveness of multiple choice options, and (4) consistency with source dataset annotations. Disagreements were resolved through majority voting, with persistent ambiguities leading to question removal.

Despite carefully selecting well-annotated datasets for construction, the validation process revealed that certain task types were still particularly prone to systemic issues. Position questions frequently exhibited ambiguous or underspecified spatial relationships, often stemming from incomplete or inconsistent relationship annotations in the source datasets. Similarly, Counting questions continued to suffer from missed instances or miscounted objects, indicating that even high-quality datasets contain non-trivial annotation noise for fine-grained quantitative reasoning. After validation and filtering, approximately 20,000 questions remained as the verified question pool.

To ensure fair and comprehensive evaluation across all dimensions, we applied stratified sampling to construct the final benchmark with balanced distribution across three key axes: Task Types, Reasoning Patterns, and Adversarial Pressures. This sampling procedure resulted in the final benchmark comprising 3,484 questions across 11,732 images, ensuring both high-quality ground truth through rigorous validation and fair evaluation coverage across all benchmark dimensions.

B. Detailed Performance Analysis

In this section, we present a comprehensive performance analysis of various multimodal large language models (MLLMs). Before delving into the complex interactions between different factors, we first examine the aggregated performance of models across three primary dimensions: **Task**, **Reasoning Pattern**, and **Adversarial Pressure.** Fig. I illustrates the performance overview. This visualization allows for a direct comparison of how different models handle specific types of challenges independently.

B.1. Variation Analysis

Fig. II presents marginal performance distributions across each dimension, illustrating their capacity to distinguish model capabilities. Among task types, Position tasks demonstrate the highest variance ($\sigma = 0.163$), suggesting that models exhibit notably different levels of spatial understanding ability. In contrast, Counting tasks show the

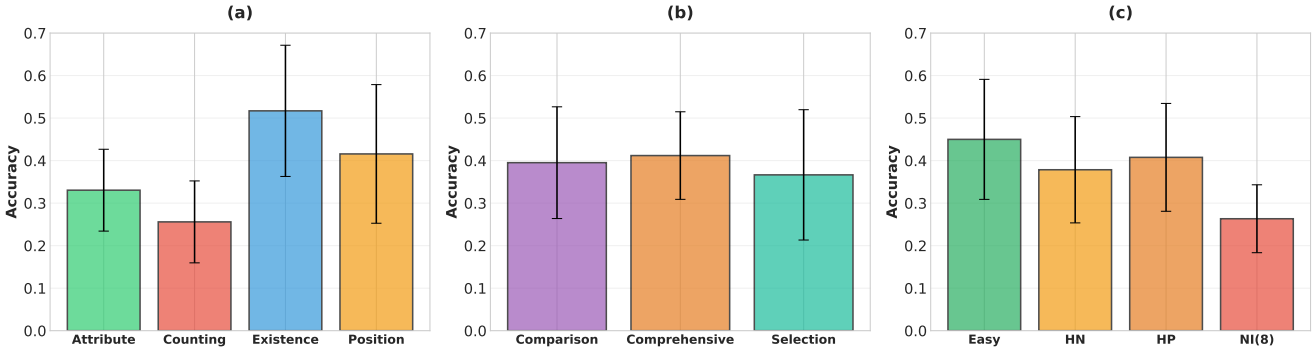


Figure II. **Performance mean and variance.** (a) Tasks, (b) Reasoning patterns, and (c) Adversarial Pressures.

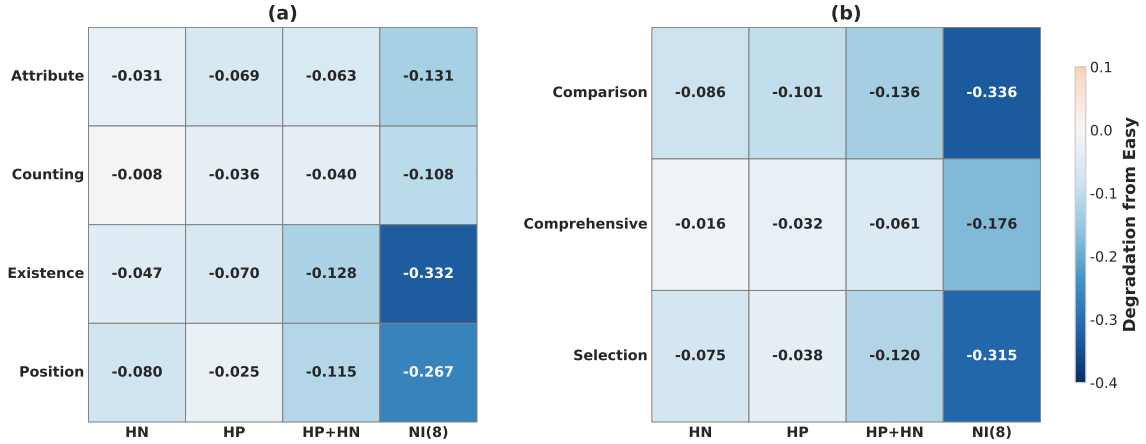


Figure III. **Cross-dimensional degradation analysis.** Performance degradation from Easy baseline across (a) Task types and (b) Reasoning dimensions under adversarial pressure.

lowest variance ($\sigma = 0.096$), indicating that this task type presents consistent difficulty across most models. Similarly, among adversarial pressure conditions, increasing Number of Images(NI) exhibits the lowest variance ($\sigma = 0.079$), comparable to Counting tasks, suggesting that extreme context pressure leads to uniformly poor performance across models. For reasoning patterns, Selection demonstrates the highest variance ($\sigma = 0.153$), suggesting that the ability to accurately identify a specific image among multiple candidates varies considerably across different models.

B.2. Cross-dimensional Interaction Analysis

Fig. III presents degradation heatmaps showing performance drops from the Easy baseline across different task-pressure and reasoning-pressure combinations, averaged over all models. Increasing Number of Images(NI) causes catastrophic degradation across all dimensions, with Comparison reasoning and Existence tasks most severely affected. Combined(HP+HN) pressure shows additive difficulty, causing larger drops than either pressure alone. Comprehensive reasoning demonstrates superior robustness compared to Comparison and Selection, suggesting holistic reasoning strategies better withstand adversarial pressure. Counting tasks show minimal degradation not due to robustness but floor effects, as their Easy baseline is already low. These patterns reveal that adversarial robustness is highly dimension-dependent.

B.3. Performance Gap Between Open-Source and Commercial Models

We compare the capabilities of leading commercial models and top-performing open-source models in Fig. IV. The performance gap between these two categories is most pronounced in Counting and Comparison tasks. These capabilities represent the areas where commercial models demonstrate the largest advantages over open-source alternatives, indicating differences in multi-image reasoning capacity.

C. Benchmark Examples

C.1. Qualitative Examples from MIOH benchmark

Figs. VI and VII provide qualitative examples from the MIOH benchmark, illustrating how questions are formulated across our four core tasks and various visual adversarial pressures. Each example is designed to test a specific

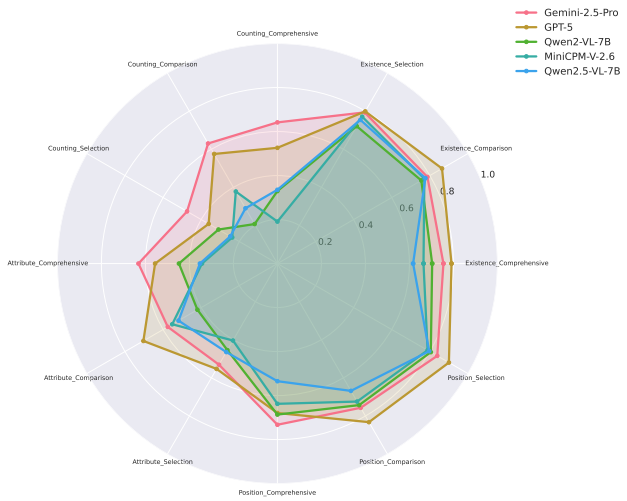


Figure IV. **Performance of commercial and top open-source models.**

aspect of an MLLM's object-centric capabilities and robustness against perceptual challenges.

Existence Tasks. (Fig. VI, top section) assess the model's ability to verify the presence of objects and pinpoint their location within the image set. The *Selective* questions showcased here require the model to identify the specific image containing the target object among candidates. Examples progress from straightforward cases (Easy: clearly visible "donut") to perceptually challenging scenarios. The hard positive example requires detecting a "bench" in a rainy scene, where the object is situated near the greenery and partially obscured by a person with umbrella, making it difficult to spot. The Hard Negative example tests robustness against contextual bias, such as identifying a "keyboard" in a set of images containing office-like environments or other electronics (e.g., game consoles) that visually resemble the target context but lack the specific object.

Counting Tasks. (Fig. VI, bottom section) evaluate quantitative reasoning capabilities through *Comprehensive* questions that require aggregating counts across all provided images. The Easy examples involve basic enumeration, such as counting the occurrences of clearly visible "elephant"s across four frames. The Hard Positive scenario challenges the model to sum the total number of "sandwiches" across images where objects are heavily occluded, cut into pieces, or cluttered on plates, testing the ability to handle dense visual information. The Hard Negative example asks for the count of "trains," where models must distinguish the actual object from contextually similar background elements like tracks or station platforms in the non-target images.

Position Tasks.(Fig. VII, left section) present the most complex spatial relationship challenges. We use *Selective* questions to ask the model to identify the image where a specific relation holds. Examples include Easy scenarios ("dog next to cat"), Hard Negative cases ("chair positioning relative to dog"), and Hard Positive examples ("person next to umbrella") that test compositional scene understanding beyond simple object detection.

Attribute Tasks. (Fig. VII, right section) assess detailed compositional understanding by requiring models to bind visual properties with objects using *Comparative* questions. The tasks range from detecting visually distinct attributes (Easy: "red scarf") to more subtle distinctions. The Hard Negative scenario involves identifying a "dark yellow mug" in a cluttered indoor environment where lighting and other objects may create confusion. The Hard Positive example tests the model's ability to consistently recognize a "white bench" across two different scenes despite variations in perspective and background.

Each example demonstrates the three question types designed for multifaceted evaluation: comprehensive (collective understanding across images), comparative (identifying differences between images), and selective (retrieving specific images matching descriptions). The progression of difficulty incorporates both visual factors (scale, occlusion, contextual bias) as detailed in Sec. 3.3.

C.2. Failure Case Examples of Commercial Models

To better understand the limitations of commercial MLLMs, we analyze specific failure cases of GPT-5 and Gemini-2.5-Pro on the MIOH benchmark. Despite their strong baseline performance, these models still exhibit hallucination patterns under visual adversarial pressures. Fig. V illustrates representative error cases where models fail to ground visual evidence correctly within multi-image contexts.

For example, in the **Attribute-Comprehensive** task, models exhibit perceptual blindness or aggregation failures when objects are visually ambiguous or occluded; for instance, both models fail to confirm the presence of "dark gray bowls" in all target images, demonstrating a breakdown in comprehensive verification. The example in **Counting-Comparative** task reveals weaknesses in fine-grained classification, where models confuse "wine glasses" with perceptually similar objects like water goblets or tumblers, leading to incorrect frequency comparisons. Most critically, the **Position-Selective** task exposes severe object and relationship hallucination under the "None of the above" option. When asked for a "cow behind a dog," models are forced into making a selection, resulting in hallucinated object identities (e.g., misidentifying a zebra as a cow in Image 4) or ignoring spatial constraints (e.g., selecting a dog-cow pair with incorrect positioning in Image 3), rather than correctly abstaining.

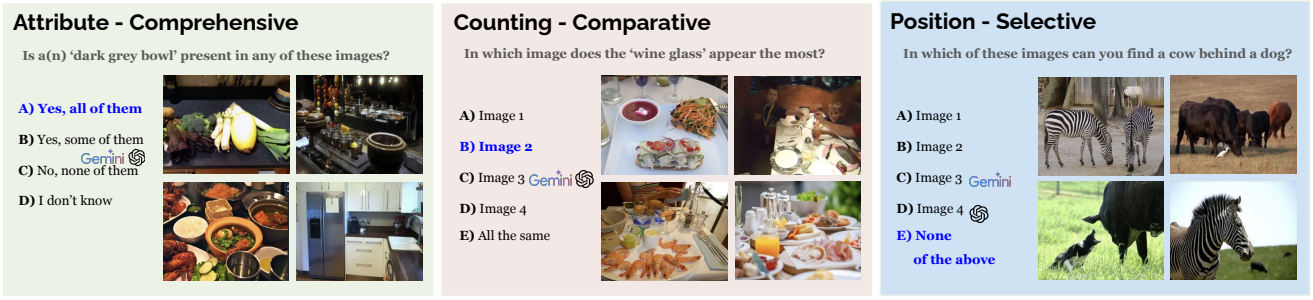


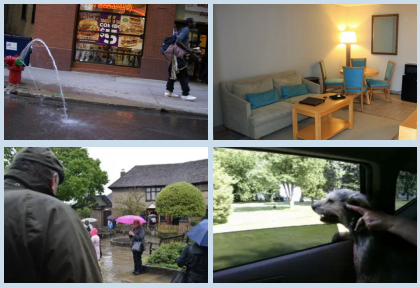
Figure V. **Failure cases of top commercial models** (GPT-5, Gemini-2.5-pro). Blue text indicates the Ground Truth (GT). The icons indicate the incorrect choices made by each model. These examples highlight how perceptual difficulty lead to reasoning failures.

Existence - Selective

EASY

In which image does a 'donut' present?

- ✓ **A) Image 1**
- B) Image 2
- C) Image 3
- D) Image 4
- E) None of the above



Hard Positive

In which image does a 'bench' present?

- A) Image 1
- B) Image 2
- ✓ **C) Image 3**
- D) Image 4
- E) None of the above

Hard Negative

In which image does a 'donut' present?

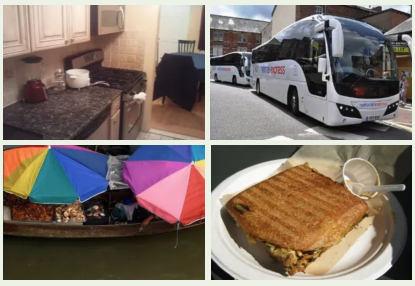
- A) Image 1
- ✓ **B) Image 2**
- C) Image 3
- D) Image 4
- E) None of the above

Hard Negative

In which image does a 'keyboard' present?

- A) Image 1
- B) Image 2
- C) Image 3
- ✓ **D) Image 4**
- E) None of the above

Counting - Comprehensive



EASY

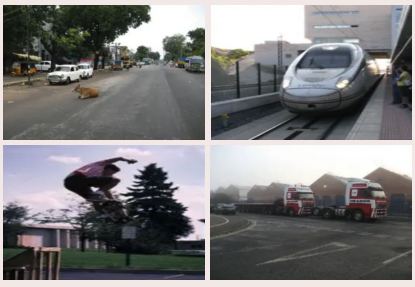
Which of the following objects appears a total of 1 time across all four images?

- A) hot dog
- B) toaster
- ✓ **C) cup**
- D) broccoli
- E) None of the above

EASY

In how many of these four images is a(n) 'elephant' present?

- A) 2 Images
- B) 3 Images
- C) 0 Image
- ✓ **D) 4 Images**
- E) I don't know



Hard Negative

In how many of these four images is a 'train' present?

- ✓ **A) 1 Image**
- B) 2 Images
- C) 3 Images
- D) 0 Image
- E) I don't know

Hard Positive

What is the total number of 'sandwiches' across all four images?

- A) 11
- B) 15
- ✓ **C) 10**
- D) 8
- E) None of the above

Figure VI. **Benchmark Examples 1.** Existence and Counting Task

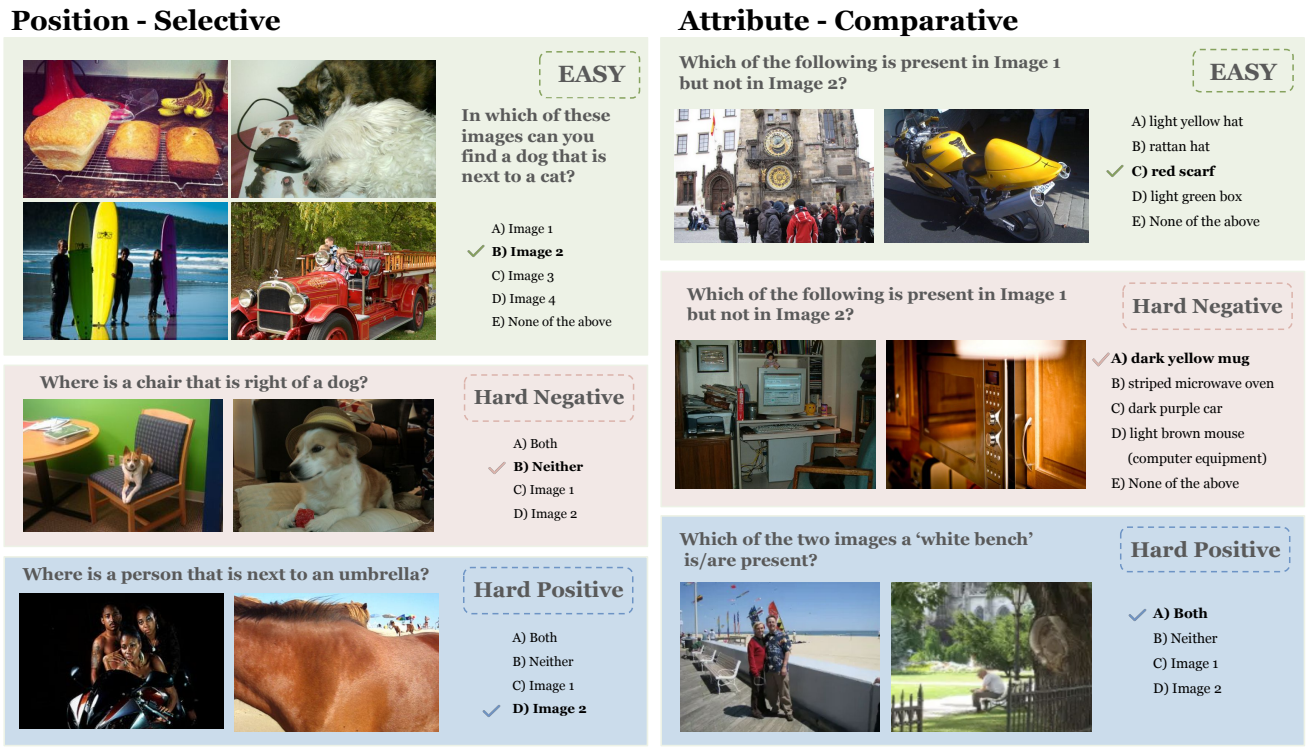


Figure VII. **Benchmark Examples 2.** Position and Attribute Task