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.

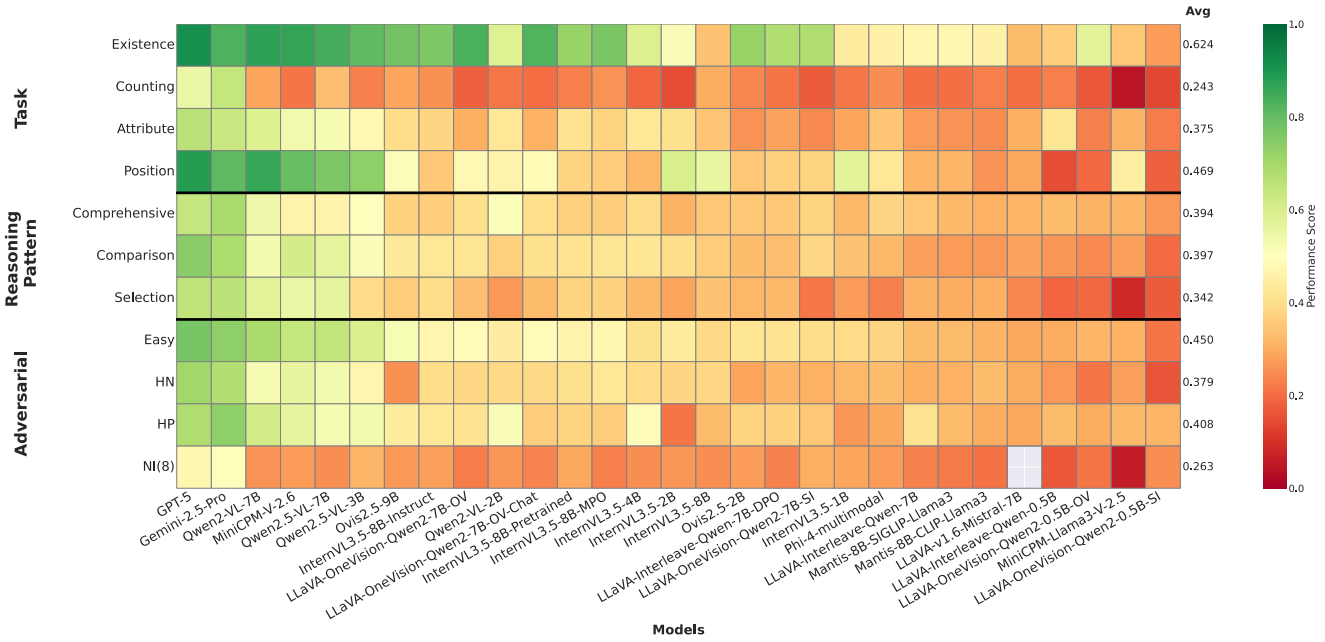


Figure I. **Aggregated Performance Heatmap across Task, Reasoning Pattern, and Adversarial Pressure Dimensions.**

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

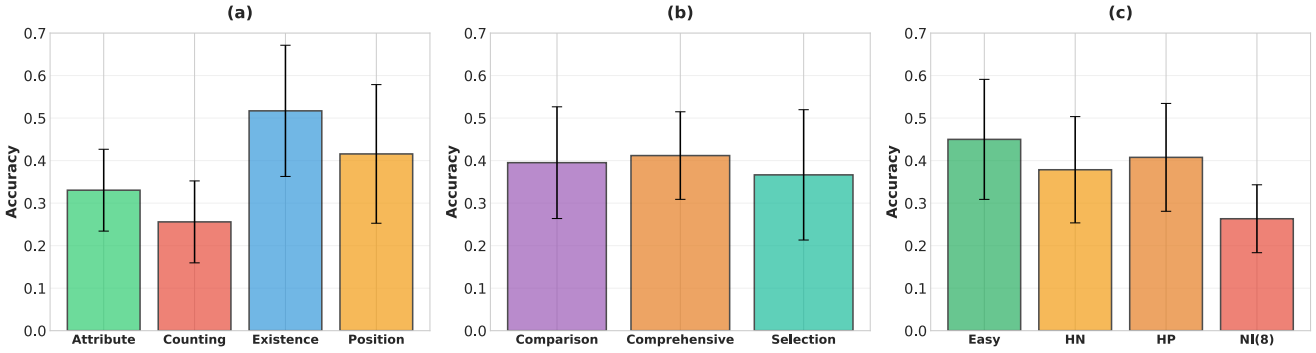


Figure II. **Performance mean and variance.** (a) Tasks, (b) Reasoning patterns, and (c) Adversarial Pressures.

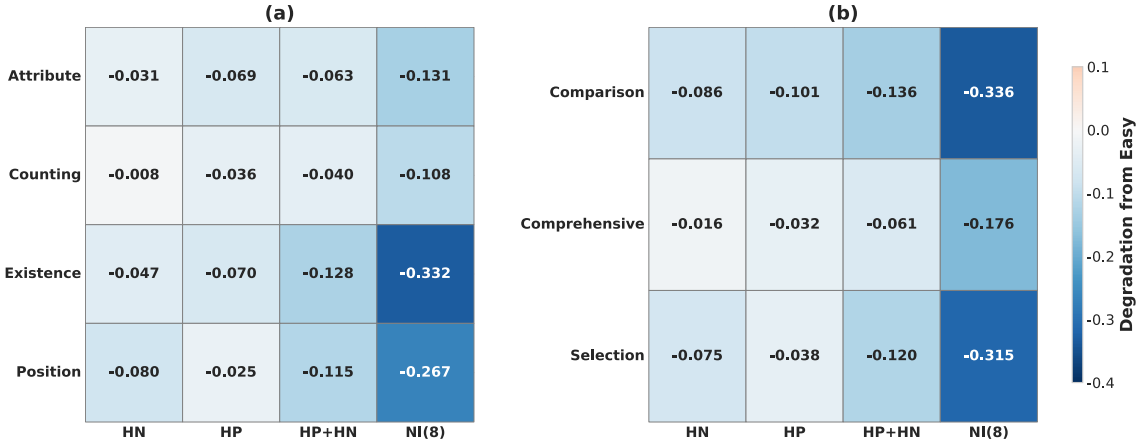


Figure III. **Cross-dimensional degradation analysis.** Performance degradation from Easy baseline across (a) Task types and (b) Reasoning dimensions under adversarial pressure.

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

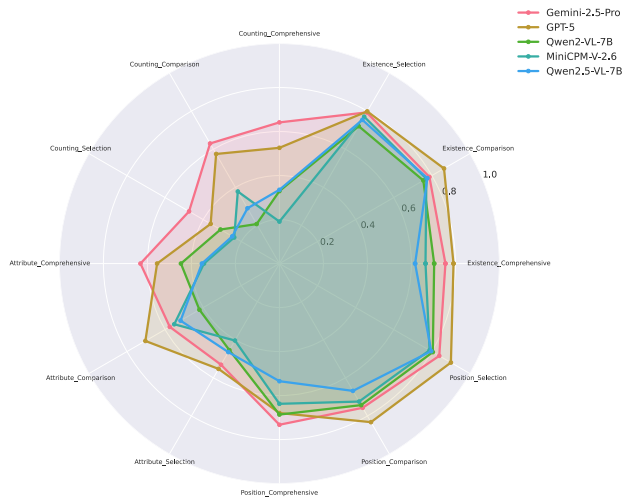


Figure IV. **Performance of commercial and top open-source models.**

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.

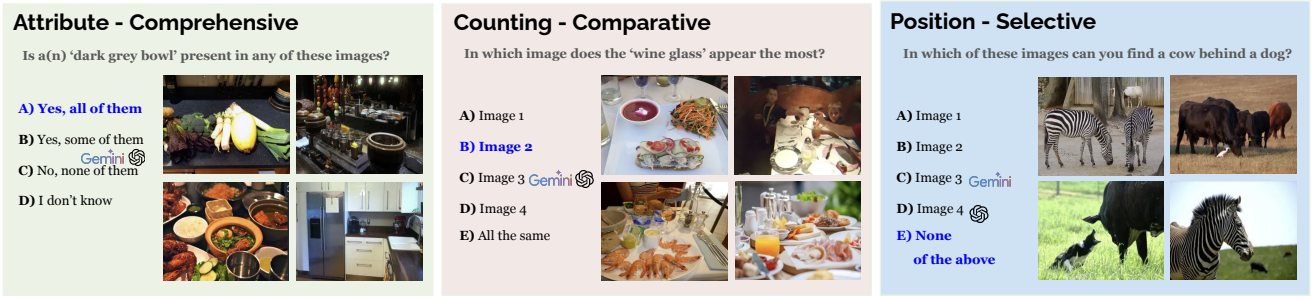


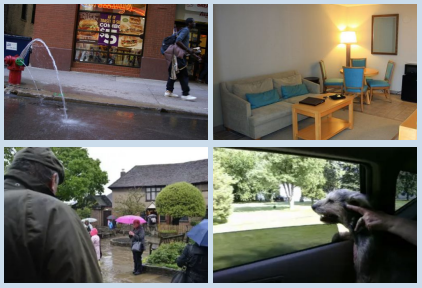
Figure V. **Failure cases of top commercial models** (GPT-5, Gemini-2.5-pro). Blue text indicates the Ground Truth (GT). The icons indicate the incorrect choices made by each model. These examples highlight how perceptual difficulty lead to reasoning failures.

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?

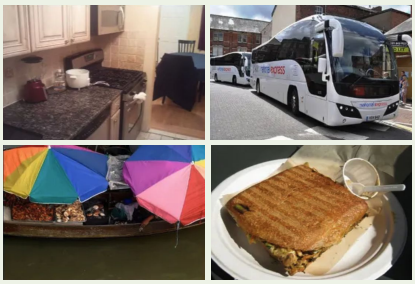
- A) Image 1
- ✓ **B) Image 2**
- C) Image 3
- D) Image 4
- E) None of the above

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

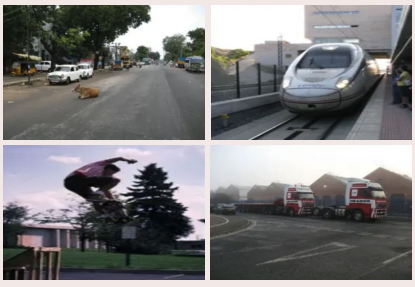
Which of the following objects appears a total of 1 time across all four images?

- 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

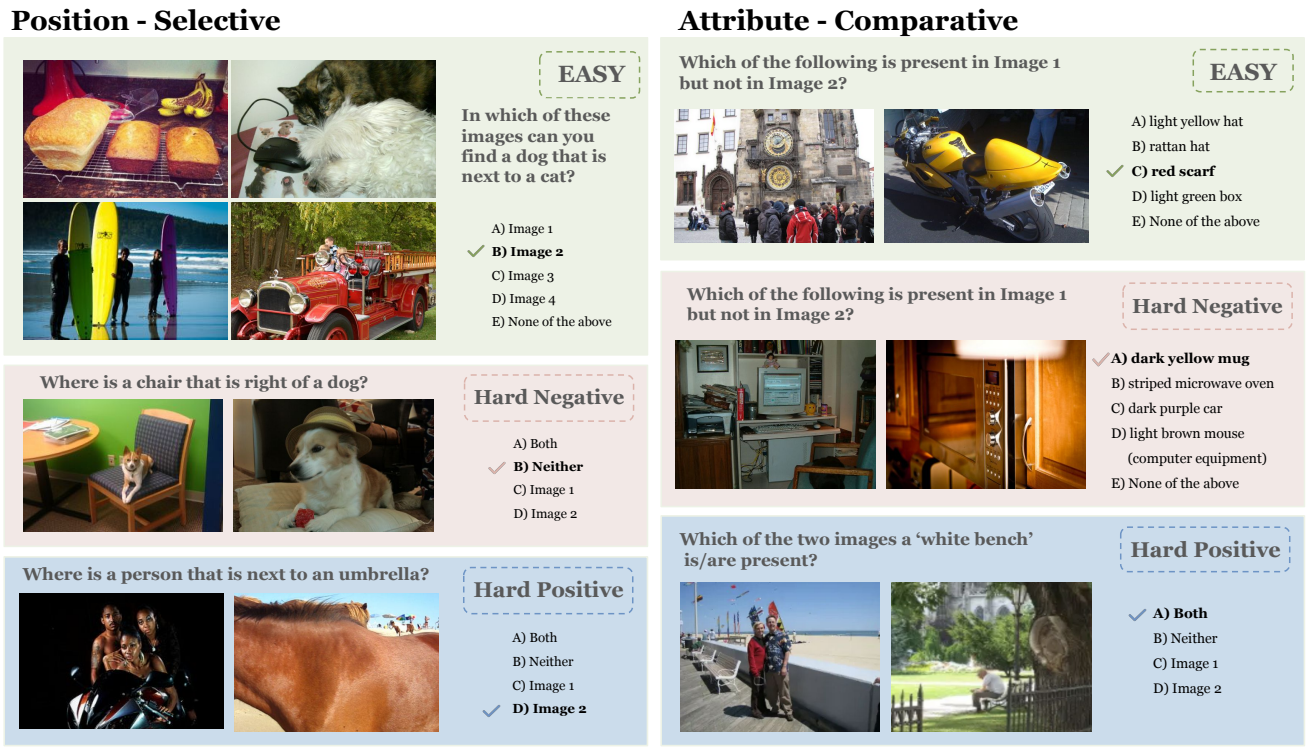


Figure VII. **Benchmark Examples 2.** Position and Attribute Task