Supplementary for GHOST: Getting to the Bottom of Hallucinations with A Multi-round Consistency Benchmark

Vibashan VS[1,2] Nadine Chang[2] Jenny Schmalfuss[2,3]
Vishal M. Patel[1] Zhiding Yu[2] Jose M. Alvarez[2]
[1]Johns Hopkins University [2]NVIDIA [3]University of Stuttgart

Introduction

As part of the supplementary materials for this paper, we provide detailed implementation information, additional experiments and analyses, an overview of the GHOST dataset, and examples from the GHOST dataset. Furthermore, we present more visual and quantitative results that extend to the ones shown in the main paper.

A. Implementation Details

A.1. Hardware information

All experiments are run on a machine with an 80GB NVIDIA A100 GPUs.

A.2. Package Details

To ensure reproducibility and efficiency, we utilized the open-source `VLMEvalKit` [6] for all experiments. The `VLMEvalKit` provides a comprehensive framework for evaluating vision-language models, streamlining the experimental workflow. Below, we outline the key libraries and parameter settings used in our implementation:

- **VLMEvalKit:** We utilized the official open-source implementation, available at https://github.com/open-compass/VLMEvalKit.
- **PyTorch:** All experiments were conducted using PyTorch version `2.1.2`.
- **Hugging Face Transformers:** We leveraged open-source MLLMs hosted on Hugging Face. Following the `VLMEvalKit` documentation, we used the corresponding Transformer packages for the respective models.
- **NumPy:** Version `1.21.2` was employed for efficient numerical computations, particularly for array and matrix operations.
- **Matplotlib** and **Seaborn:** For data visualization, we used Matplotlib version `3.4.3` and Seaborn version `0.11.2`.

A.3. Models

We detail the sources of the pretrained models we use in the paper for evaluation.

- **GPT-4o**: The API key can be found at https://platform.openai.com/docs/api-reference/introduction.
- **Gemini 1.5 Pro**: The API key can be found at https://ai.google.dev/gemini-api/docs?gad_source=1&gclid=Cj0KCQiA6Ou5BhCrARIsAPoTxrBNiYImqyaX_VkK97FSSAtzQ2nAfN__Ksk_hJut5SdzpVxrWvVOsBQaAqpVEALw_wcB.
- **LLaVA-OneVision Qwen2 0.5B**: The model can be accessed here: https://huggingface.co/llava-hf/llava-onevision-qwen2-0.5b-ov-hf.
- **LLaVA-OneVision Qwen2 7B OV**: The pretrained model is available at https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov.
- **LLaVA-OneVision Qwen2 72B OV**: The pretrained model can be accessed here: https://huggingface.co/lmms-lab/llava-onevision-qwen2-72b-ov.
- **LLaVA v1.5 13B**: We obtained the pretrained model released by its author at https://huggingface.co/liuhaotian/llava-v1.5-13b.
- **Chameleon 7B**: The model is provided by Meta and available at https://huggingface.co/facebook/chameleon-7b.
- **CogVLM2 LLaMA3 Chat 19B**: The pretrained model is accessible at https://huggingface.co/THUDM/cogvlm2-llama3-chat-19B.
- **IDEFICS 9B**: The model is available at https://huggingface.co/HuggingFaceM4/idefics-9b-instruct.
- **MiniCPM-V**: The pretrained model is provided at https://huggingface.co/openbmb/MiniCPM-V.

Model	Size	Vision Encoder	LLM
LLaVA-OneVision [13]	0.5B	SigLIP-So400m	Qwen2-0.5b
MiniCPM-V [18]	2B	SigLIP-So400m	MiniCPM-2.4B
VILA 1.5 [14]	3B	SigLIP-So400m	Sheared LLaMA-2.7b
PaliGemma [3]	3B	SigLIP-So400m	Gemma-2B
Phi-3.5-Vision [2]	4B	CLIP ViT-L/14	phi-3.5-mini
Mantis-LLaMA3 [10]	8B	SigLIP-So400m	LLama3-8b-instruct
VILA 1.5 [14]	13B	SigLIP-So400m	LLama3-8b-instruct
Eagle-X4-Plus [16]	8B	Mixture of Vision Encoder	LLama3-8b-instruct
LLaVA-OneVision [13]	7B	SigLIP-So400m	Qwen2-7b
Idefics-9b [12]	7B	CLIP-ViT-H-14-laion2B-s32B-b79K	LLama-7b
LLaVA-v1.5 [15]	7B	CLIP ViT-L-14	Vicuna-7b
VILA 1.5 [14]	13B	SigLIP-So400m	Vicuna-13b
MiniCPM-Llama3 [8]	18B	SigLIP-So400m	Llama3-8B-Instruct
CogVLM2-19b [7]	20B	EVA-CLIP	Llama-3-8B-Instruct
LLaVA-v1.5 [15]	13B	CLIP ViT-L-14	Vicuna-13b
InternVL-Chat [5]	26B	InternViT-6B	InternLM2-Chat
VILA 1.5 [14]	40B	SigLIP-So400m	Yi-34B
LLaVA-OneVision [13]	72B	SigLIP-So400m	Qwen2-72B

Table A.1. Multi-modal LLM meta-data about total size, vision encoder and LLM.

- **MiniCPM-Llama3-V-2.5**: The pretrained model can be found here: https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5.
- **Phi-3.5 Vision Instruct**: The model is provided by Microsoft and available at https://huggingface.co/microsoft/Phi-3.5-vision-instruct.
- **VILA1.5-3B**: The model can be found here: https://huggingface.co/Efficient-Large-Model/VILA1.5-3b.
- **VILA1.5-8B**: The pretrained model is available at https://huggingface.co/Efficient-Large-Model/Llama-3-VILA1.5-8B.
- **VILA1.5-13B**: The model is accessible at https://huggingface.co/Efficient-Large-Model/VILA1.5-13b.
- **VILA1.5-40B**: The pretrained model is provided at https://huggingface.co/Efficient-Large-Model/VILA1.5-40b.
- **InternVL-Chat**: The model is accessible at https://huggingface.co/OpenGVLab/InternVL-Chat-V1-5.
- **PaliGemma 3B Mix 448**: The model is available at https://huggingface.co/google/paligemma-3b-mix-448.
- **Eagle-X4**: The model is accessible at https://hf.rst.im/NVEagle/Eagle-X4-8B-Plus.

B. Additional Discussion and Experiments

Table B.1. Compositional Triplet Comparison

Object	OAR Question	Type	CHAIR	POPE	MHaluBench	AMBER	GHOST
Lamp	is there lamp?	exist	✓	✓	✓	✓	✓
	is the lamp red?	attr	✗	✗	✓	✓	✓
	is lamp on table?	rel	✗	✗	✗	✗	✓
Table	is there table?	exist	✓	✓	✗	✓	✓
	is the table wooden?	attr	✗	✗	✓	✗	✓
	is table under lamp?	rel	✗	✗	✗	✗	✓

B.1. Comparison Between Object-Centric and Image-Level Benchmarks

While existing benchmarks (e.g., MHaluBench [4], AMBER [17]) do include fine-grained visual concepts, they predominantly operate at an *image-level*, meaning they do not systematically ensure coverage of each object's existence, attributes, and relationships. In contrast, **GHOST** adopts an *object-centric* perspective, comprehensively evaluating these three dimensions (Exist, Attr, Rel) for *every* object in a scene.

Table B.1 illustrates this difference by comparing representative queries across benchmarks for two objects, *lamp* and *table*. Each row demonstrates whether a benchmark addresses specific object-oriented questions about

existence (e.g., “is there a lamp?”), attributes (e.g., “is the lamp red?”), or relations (e.g., “is the lamp on the table?”). Unlike other benchmarks, GHOST systematically covers all three dimensions for each object and additionally introduces robust negative samples for each dimension. This approach ensures a thorough, itemized check of a model’s ability to handle object-level understanding and reasoning, rather than focusing solely on image-level content.

B.2. GCS Calibration.

In order to assess the robustness of the proposed GHOST Consistency Score (GCS), we conducted a calibration study by varying the weights used in the weighted geometric mean computation. Specifically, we experimented with weights of 1, 1.5, 2, and 3. Although the absolute GCS values shifted with these different weight settings, the relative ranking of the models remained consistent (see Table B.2). This finding suggests that even low-severity hallucinations—while affecting the overall score magnitude—do not disproportionately distort the model rankings. In other words, the calibration process confirms that the GCS is stable and that its ability to capture a model’s tendency to hallucinate is not overly sensitive to the specific choice of weighting parameters.

Table B.2. GCS Scores for Different Weighting Schemes

Model	GCS (2)	GCS (1)	GCS (1.5)	GCS (3)
Phi-3.5-V	55.8	65.6	62.4	48.6
LLaVA-1.5-13B	58.2	67.6	64.6	51.1
MiniCPM-LLaMA3	63.9	72.3	70.0	57.1
VILA1.5-13b	64.5	72.6	70.2	58.0
GPT-4o	69.0	76.4	74.5	62.8

B.3. Hallucination Benchmarks Comparison

We examine the performance of various models on popular MLLM benchmarks and our benchmark, see Tab. B.3. While VILA 1.5 13B demonstrates weaker performance than Phi-3-vision on MMbench (74.9 vs 81.9) and MMMU (val) (37.9 vs 43.0), it surpasses Phi-3-vision in terms of hallucination performance on GHOST (64.5 vs. 55.8). This suggests that proficiency in visual question answering and reasoning tasks does not necessarily equate to better performance in reducing hallucinations. On the other hand, GPT-4o emerges as the overall top model with the best scores in MMbench, MMMU and GHOST, which highlights its comprehensive strength across benchmarks. However, there is significant potential for further improvement in mitigating hallucinations, even among recent SOTA MLLMs.

B.4. Analysis of "True/False" Bias and Evaluation Metrics

To determine whether our results might be influenced by a systematic bias toward answering “False,” we conducted a theoretical analysis of how such a bias would affect the GHOST Consistency Score. Given a probability p of answering “False” to any question, we calculated the expected consistency score using the geometric weighting formula:

$$\text{Score}(p) = 1 - \frac{\mathbb{E}[\text{incorrect}] \cdot \left(1 + \frac{1}{2} + \frac{1}{4}\right)}{1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}}, \tag{1}$$

where the expected number of incorrect answers is given by

$$\mathbb{E}[\text{incorrect}] = p + 3 \cdot (1 - p).$$

The results, as shown in Table B.4, indicate that even a model that always answers “False” (i.e., $p = 1$) achieves a consistency score of 46.7%. This baseline score arises due to the structure of the GHOST Consistency Scoring system and reflects the inherent penalties when the model fails to correctly answer the questions.

In our evaluation setup, each test consists of four true/false questions with one true statement and three false statements. Consequently, even a single hallucination (i.e., an incorrect prediction on the true statement) is penalized heavily. To further characterize model performance, we also computed the following metrics:

- **Accuracy:** The ratio of correct answers (true positives and true negatives) to the total number of questions.
- **Precision (for the True class):** The proportion of correctly predicted true answers among all questions predicted as true.
- **Recall (for the True class):** The proportion of actual true instances that are correctly identified.

Model	MME	MMBench	MMMU	GHOST
Phi-3.5-V [2]	-	81.9	43.0	55.8
LLaVA-1.5 13B [15]	-	67.8	36.4	58.2
VILA 1.5 13B [14]	1569.5	74.9	37.9	64.5
MiniCPM-Llama3 [8]	2024.6	77.2	45.8	63.9
LLaVA-OneVision-7b [13]	1580.0	80.8	48.8	64.4
GPT-4o [1]	-	**83.4**	**69.1**	**69.0**

Table B.3. Models on popular MLLM benchmarks and proposed GHOST benchmark.

False Probability (p)	Expected Incorrect	Consistency Score (%)	Accuracy (%)	Precision (%)	Recall (%)
0.0	3.000	6.7	25.0	25.0	100.0
0.3	2.100	22.7	40.0	25.0	70.0
0.5	1.500	33.3	50.0	25.0	50.0
0.7	0.900	44.0	60.0	25.0	30.0
1.0	1.000	46.7	75.0	0.0	0.0

Table B.4. Theoretical GHOST Consistency Scores and Performance Metrics Under Different "False" Response Probabilities.

For a given probability p of answering "False," we derive the following:

$$\begin{aligned} \text{True Positives (TP)} &= 1-p, && \text{(for the sole true question)} \\ \text{False Negatives (FN)} &= p, \\ \text{True Negatives (TN)} &= 3p, && \text{(for the three false questions)} \\ \text{False Positives (FP)} &= 3(1-p). \end{aligned}$$

Thus, the metrics can be computed as:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{4} = \frac{(1-p)+3p}{4} = \frac{1+2p}{4},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{1-p}{1-p+3(1-p)} = \frac{1-p}{4(1-p)} = 0.25,$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} = 1-p.$$

For the edge case $p = 1$ (i.e., when the model always answers "False"), the model misses the sole true instance, resulting in a precision of 0% for the true class. Table B.4 summarizes the theoretical values for the consistency score, accuracy, precision, and recall across different values of p. In this scenario, a model that always responds "False" achieves a biased baseline score solely due to the inherent penalty structure—this score arises from the weighting of incorrect responses, even though the precision and recall for the true instance are null.

This analysis shows that the metrics— GCS, accuracy, precision, and recall—are not mere artifacts of a response bias. Instead, they reflect the model's ability to correctly identify and represent factual details regarding objects, attributes, and relations. This confirms from our main experimental table that their low GHOST Consistency Scores, accuracy, precision, or recall stem from genuine hallucination issues in object-level, attribute, and relational understanding rather than from any bias toward a particular binary answer.

B.5. Text-only Experiment

To illustrate our benchmark requires both text and images, we conducted a text-only experiment where the models received only textual descriptions without any visual input. This approach ensures that there is no data leakage from the text alone. The results in Table B.5 show that text-only performance is limited across all categories, with the models consistently achieving similar scores around 46.7%. This performance aligns with the expected score for a model that defaults to answering "False" due to the lack of visual input, as discussed in Section B.4. This uniform performance indicates that without images, the models cannot accurately interpret the tasks. Therefore, the results confirm there is no data leakage from the textual modality alone, emphasizing the benchmark's robustness in evaluating multimodal understanding.

B.6. Limitations

While GHOST allows for more detailed hallucination analysis and evaluation for MLLMs, it has a few limitations.

Models	Obj	Attr	Rel	Overall (%)
Phi-3.5-V	78.3	52.2	36.9	55.8
Phi-3.5-V (text-only)	46.4	47.9	44.4	46.2
LLaVA-OneVision-7B-OV	86.5	67.5	39.1	64.4
LLaVA-OneVision-7B-OV (text-only)	46.6	45.9	45.2	45.9
MiniGPT-4-Llama2-7B	79.9	65.4	46.4	63.9
MiniGPT-4-Llama2-7B (text-only)	44.8	46.7	47.0	46.1

Table B.5. MLLM performance for the text-only experiment.

1. We construct our conditional negatives in GHOST by using frequent word co-occurrences seen in the LLAVA instruct dataset. However, the full language vocabulary is vast, and negatives constructed from a broader world knowledge vocabulary would be more ideal and representative.
2. GHOST is built on the GQA dataset and scene graphs, where images contain older object representations of the world (e.g., flip phones). Furthermore, the captions are generally shorter and simpler than the complex captions MLLMs process and output today.
3. Recently, many MLLMs have started including new data in their training datasets. GQA is one of the highest-quality datasets that provide objects, attributes, and relations. If an MLLM is trained on these GQA scene graphs, there might be data leakage, potentially improving performance. However, it is worth noting that our key contribution is evaluating hallucination using consistency checks at the object level, which helps to understand whether a model truly understands objects in an image.
4. In the future, we plan to extend this benchmark based on the GHOST idea, creating a closed evaluation protocol. Instead of using existing GQA images and LLaVa vocabulary with object, attribute, and relation annotations, we will manually curate positive and negative pairs with images with no text annotation, such as SAM, to evaluate MLLMs.
5. Coverage is bounded by Visual Genome/GQA distributions and our manual hard-negative design.
6. GHOST targets object-level hallucination; higher-level temporal/commonsense reasoning is out of scope in this release.

C. GHOST Dataset Details

We obtain all existing datasets from their original sources released by the authors. We refer readers to these sources for the dataset licenses. To the best of our knowledge, the data we use does not contain personally identifiable information or offensive content.

GHOST data For GHOST, we utilize Visual Genome dataset [11] and GQA dataset [9] and images from GQA val set.

- **Visual Genome:** The visual genome dataset, paper and additional details [1].
- **GQA:** The GQA dataset, paper and additional details [2].

Co-occurrence data For conditional negative sampling we utilized Visual Genome, VQA, GQA, OCR VQA and LLAVA instruction tuning datasets.

- **Visual Genome:** The visual genome dataset, paper and additional details [3].
- **VQA:** The VQA dataset, paper and additional details [4].
- **OCR VQA:** The OCR VQA dataset, paper and additional details [5].
- **GQA:** The GQA dataset, paper and additional details [6].
- **Text VQA:** The Text VQA dataset, paper and additional details [7].

C.1. GUI configuration

As shown in Figure C.1, the primary task of the GUI is to assist in accurately identifying and categorizing images, with a focus on verifying 'hard negatives'—data points that are incorrect or unrelated synonyms of positive labels. This is crucial for refining the dataset used in training machine learning models. The GUI's functionality is essential

[1] https://homes.cs.washington.edu/~ranjay/visualgenome/index.html
[2] https://cs.stanford.edu/people/dorarad/gqa/about.html
[3] https://homes.cs.washington.edu/~ranjay/visualgenome/index.html
[4] https://visualqa.org/
[5] https://ocr-vqa.github.io/
[6] https://cs.stanford.edu/people/dorarad/gqa/about.html
[7] https://textvqa.org/

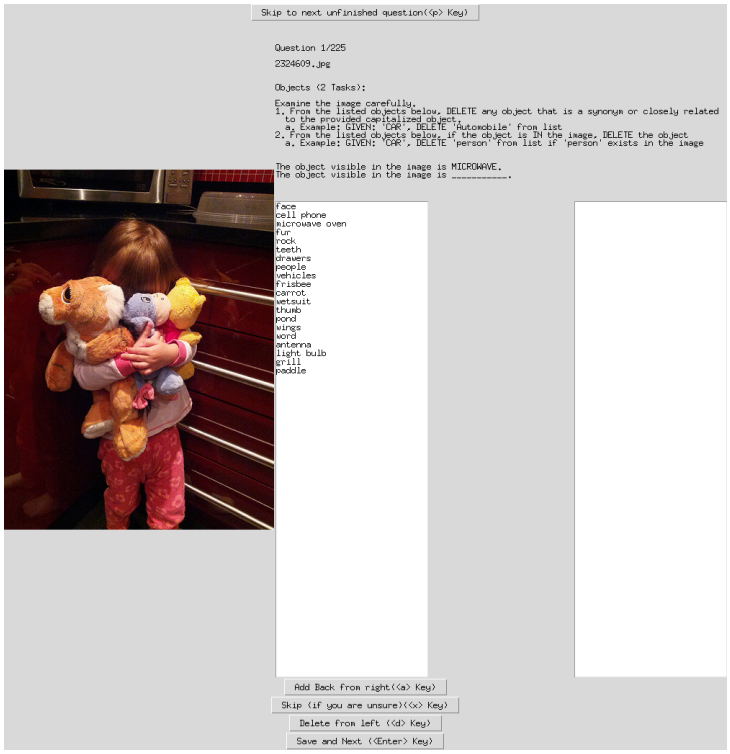

Figure C.1. GUI to remove false positives to generate high-quality conditional negatives

for ensuring the integrity of data by allowing precise verification and removal of these inaccuracies, thus enhancing model reliability and performance. The GUI is developed using Python where it employs the Pillow library for image processing and NumPy for efficient data management, facilitating robust image data and annotation handling. The interface features a dual-pane layout, with the image on the left and interactable labels on the right, designed for streamlined navigation and annotation. Keyboard shortcuts enable easy progression through images and labels, significantly reducing annotator fatigue and increasing productivity. This setup is particularly effective for large-scale image datasets, ensuring thorough manual review and optimization of data filtering.

C.2. Filtering Instruction

The filtering process within the GUI is divided into three main tasks: Objects, Attributes, and Relations, each with specific guidelines:

Objects Tasks

1. **Synonym Filtering:** Examine each listed object carefully. If an object is a synonym or closely related to the provided capitalized reference object, it should be deleted from the list. For example, if `"CAR"` is given, you would delete `"Automobile"` from the list.
2. **Existence Verification:** Check whether the object mentioned in the list physically appears in the image. If it does, delete that object from the list. For example, if `"CAR"` is given and there is a visible `"person"` in the image, then delete `"person"` from the list.

Attribute Task

- **Attribute Synonym Filtering:** Review the image to identify any attributes listed that are synonyms or closely related to a provided capitalized attribute. Delete any that qualify. For instance, given `"Bright,"` you should delete `"Luminous"` from the list.

Relation Task

- **Relation Synonym Filtering:** Analyze the image to determine if any listed relations are synonyms or closely related to the provided capitalized relation. Delete the synonyms as appropriate. For example, if `"By"` is given, delete `"Beside"` from the list.

These structured tasks aim to streamline the data annotation process, ensuring the data's accuracy and relevance by removing redundant or incorrect labels. This careful filtration enhances the quality of the training set for more effective machine learning model development.

C.3. Annotator Recruitment, Payment, Demographics

We recruited colleagues and undergraduate and graduate students through workplace. All annotators have basic background in computer science and coding. Annotators' nationalities include but not limited to the following alphabetical list - American (US), Chinese, German, Indian, Korean, and Swiss. All annotators were given a chunk of data to label that takes approximately 3 hours to complete, but completion was not required. Because our GUI allowed users to easily pick up where they left off, annotators can and did complete labeling over a large stretch of time. All annotators were adequately paid because labeling counted towards working hours, which is paid above minimum wage. Lastly, all annotators consented to use their annotated data for research purposes.

C.4. Dataset Templates and Prediction Format

The GHOST evaluation benchmark dataset and predictions are structured to evaluate models on object and scene recognition tasks based on textual prompts, object IDs, and associated images. Each entry contains essential keys for consistent benchmarking and analysis.

C.4.1. Dataset Statement Templates

The statements in the dataset are categorized into three types, each following a default template:

- **Object Statements:** These describe the presence of an object in an image.
  Template: `"A [object] is present in the image."`
  Example: `"A stuffed bear is present in the image."`
- **Attribute Statements:** These describe an attribute of an object.
  Template: `"The [attribute] of the [object] present in the image is [value]."`
  Example: `"The color of the stuffed bear present in the image is white."`
- **Relation Statements:** These describe the relationship between two objects.
  Template: `"The [relation] between the [object1] and [object2] is that the [object1] is [description] the [object2]."`
  Example: `"The spatial relation between the stuffed bear and the table is that the stuffed bear is on top of the table."`

C.4.2. Prediction Format

The general format of a dataset entry is as follows:

```
{
    "question_id": "unique_question_id",
    "object_id": "associated_object_id",
    "image": "image_filename",
    "text": "A textual description of the object or scene.",
    "label": "ground_truth_label",
    "model_name": "evaluated_model_name",
    "prediction": "model_prediction"
}
```

C.4.3. Example Entry

Below is an example entry from the GHOST benchmark dataset:

```
{
    "question_id": "1066_obj1_1pos",
    "object_id": "1066_obj1",
    "image": "1066.jpg",
    "text": "The color of the stuffed bear present in the image is white.",
    "label": "True",
    "model_name": "GeminiPro1-5",
    "prediction": "true"
}
```

C.4.4. Explanation of Keys

- **question_id:** Unique identifier for each evaluation question. For instance, `1066_obj1_1pos` indicates the object and the question type (positive/negative).
- **object_id:** Identifier linking to a specific object within the dataset (e.g., `1066_obj1`).
- **image:** Filename of the image associated with the object (e.g., `1066.jpg`).

- **text:** A OAR statement about the object (e.g., "The color of the stuffed bear present in the image is white.").
- **label:** Ground truth label (e.g., `"True"` or `"False"`).
- **model_name:** MLLM Name (e.g., `"GeminiPro1-5"`).
- **prediction:** Prediction generated by the model for the given question (e.g., `"true"`).

GHOST Unique Objects

Object Names
‘stuffed bear’, ‘vehicles’, ‘license plate’, ‘sign’, ‘lamp’, ‘people’, ‘ceiling’, ‘monitor’, ‘car’, ‘chairs’, ‘ball’, ‘pot’, ‘doll’, ‘speaker’, ‘backpack’, ‘logo’, ‘clock’, ‘post’, ‘outlet’, ‘drapes’, ‘basket’, ‘pants’, ‘grass’, ‘window’, ‘towels’, ‘luggage’, ‘skateboard’, ‘bed’, ‘bag’, ‘trees’, ‘fire’, ‘helmet’, ‘skier’, ‘spots’, ‘zebra’, ‘bat’, ‘branch’, ‘letters’, ‘tie’, ‘giraffe’, ‘desk’, ‘bears’, ‘tree’, ‘sidewalk’, ‘can’, ‘suit’, ‘handbag’, ‘racket’, ‘tire’, ‘sandwich’, ‘glass’, ‘kite’, ‘glove’, ‘table’, ‘box’, ‘horse’, ‘wine glass’, ‘man’, ‘suitcase’, ‘phone’, ‘winter coat’, ‘giraffes’, ‘feet’, ‘wheels’, ‘fence’, ‘water’, ‘passenger’, ‘chair’, ‘truck’, ‘drink’, ‘rug’, ‘shelves’, ‘snow’, ‘flag’, ‘sink’, ‘refrigerator’, ‘cup’, ‘laptop’, ‘toilet bowl’, ‘books’, ‘shower doors’, ‘kettle’, ‘pillow’, ‘napkin’, ‘counter’, ‘cars’, ‘hand’, ‘men’, ‘towel’, ‘bottle’, ‘cream’, ‘jacket’, ‘street’, ‘picture’, ‘couch’, ‘door’, ‘bread’, ‘lady’, ‘cloud’, ‘vase’, ‘bike’, ‘fur’, ‘fireplace’, ‘containers’, ‘bathtub’, ‘microwave’, ‘field’, ‘clouds’, ‘keyboard’, ‘lid’, ‘watch’, ‘wine’, ‘wall’, ‘coffee table’, ‘vehicle’, ‘tablet’, ‘curtains’, ‘boxes’, ‘sand’, ‘lemons’, ‘paper’, ‘mountain’, ‘person’, ‘mouth’, ‘word’, ‘broken branch’, ‘building’, ‘bench’, ‘symbol’, ‘pole’, ‘kitchen’, ‘windows’, ‘sea’, ‘street sign’, ‘snowboarder’, ‘nose’, ‘finger’, ‘bush’, ‘dog’, ‘coffee’, ‘traffic light’, ‘shirt’, ‘orange’, ‘bus’, ‘letter’, ‘frisbee’, ‘airplane’, ‘sun’, ‘umbrella’, ‘aircraft’, ‘dirt’, ‘camera’, ‘elephant’, ‘jeans’, ‘plate’, ‘paper plate’, ‘fries’, ‘utensil’, ‘menu’, ‘hot dog’, ‘girl’, ‘potato’, ‘sheet’, ‘number’, ‘coffee mug’, ‘child’, ‘label’, ‘belt’, ‘teddy bear’, ‘sweater’, ‘device’, ‘vest’, ‘cone’, ‘player’, ‘skateboarder’, ‘beach’, ‘bushes’, ‘trunk’, ‘shore’, ‘game’, ‘street light’, ‘fork’, ‘baseball’, ‘bicycles’, ‘wetsuit’, ‘steps’, ‘net’, ‘zebras’, ‘wheel’, ‘headboard’, ‘step’, ‘stone’, ‘spoons’, ‘controller’, ‘pan’, ‘beverage’, ‘dish’, ‘stove’, ‘floor’, ‘cups’, ‘closet’, ‘candle’, ‘beer’, ‘head’, ‘flower’, ‘coat’, ‘computer mouse’, ‘television’, ‘fire hydrant’, ‘wires’, ‘cell phone’, ‘wings’, ‘shorts’, ‘milk’, ‘goal’, ‘numbers’, ‘mountains’, ‘bear’, ‘helmets’, ‘bowl’, ‘leaves’, ‘flowers’, ‘uniform’, ‘eye’, ‘bird’, ‘dress’, ‘train’, ‘batter’, ‘tables’, ‘console’, ‘platform’, ‘elephants’, ‘book’, ‘wine bottle’, ‘food’, ‘sanitizer’, ‘surfboard’, ‘dress shirt’, ‘paper towels’, ‘boots’, ‘roses’, ‘drawers’, ‘purse’, ‘walnuts’, ‘pots’, ‘carrot’, ‘liquid’, ‘tomato’, ‘machine’, ‘peppers’, ‘avocado’, ‘knife’, ‘face’, ‘frame’, ‘plant’, ‘toilet seat’, ‘park’, ‘banana’, ‘heart’, ‘placemat’, ‘bicycle’, ‘water bottle’, ‘scissors’, ‘candies’, ‘couple’, ‘mirror’, ‘bowls’, ‘tablecloth’, ‘baby’, ‘screen’, ‘salad’, ‘cake’, ‘calf’, ‘pen’, ‘roof’, ‘cones’, ‘house’, ‘ladder’, ‘hands’, ‘buildings’, ‘pilot’, ‘road’, ‘animals’, ‘woman’, ‘jet’, ‘bun’, ‘dishwasher’, ‘frosting’, ‘mustard’, ‘oven’, ‘eyes’, ‘ketchup’, ‘noodles’, ‘dispenser’, ‘crust’, ‘mug’, ‘tongue’, ‘cable car lines’, ‘meal’, ‘church’, ‘ears’, ‘branches’, ‘motorcycle’, ‘pitcher’, ‘stop sign’, ‘collar’, ‘legs’, ‘wing’, ‘curtain’, ‘propeller’, ‘seat’, ‘boat’, ‘animal’, ‘chain’, ‘tree trunk’, ‘tail’, ‘lake’, ‘hair’, ‘railroad’, ‘cross’, ‘ear’, ‘ocean’, ‘palm tree’, ‘flower pot’, ‘glasses’, ‘ground’, ‘faucet’, ‘boy’, ‘countertop’, ‘container’, ‘eye glasses’, ‘cabinet’, ‘pipe’, ‘skirt’, ‘crowd’, ‘feathers’, ‘cow’, ‘meat’, ‘log’, ‘hat’, ‘arm’, ‘van’, ‘jersey’, ‘shoe’, ‘plants’, ‘outfit’, ‘toilet’, ‘papers’, ‘cats’, ‘guy’, ‘card’, ‘snowboard’, ‘toppings’, ‘cat’, ‘gloves’, ‘town’, ‘costume’, ‘toilet paper’, ‘rock’, ‘carpet’, ‘urinal’, ‘pillows’, ‘toy’, ‘donuts’, ‘pickles’, ‘home plate’, ‘audience’, ‘beard’, ‘tennis ball’, ‘ornament’, ‘brush’, ‘pavement’, ‘fruit’, ‘shelf’, ‘apple’, ‘blood’, ‘cheese’, ‘lettuce’, ‘sky’, ‘bananas’, ‘guitar’, ‘mound’, ‘spectators’, ‘women’, ‘sneakers’, ‘remote control’, ‘words’, ‘sweatshirt’, ‘rice’, ‘carrots’, ‘candle holder’, ‘spinach’, ‘eggs’, ‘bridge’, ‘arms’, ‘parking meter’, ‘package’, ‘paw’, ‘leg’, ‘folding chair’, ‘tower’, ‘skin’, ‘decoration’, ‘cap’, ‘painting’, ‘light switch’, ‘teeth’, ‘tv stand’, ‘surfer’, ‘socks’, ‘blanket’, ‘goggles’, ‘umpire’, ‘scarf’, ‘pizza’, ‘display’, ‘skis’, ‘computer’, ‘fan’, ‘printer’, ‘stuffed animal’, ‘bracelet’, ‘binder’, ‘olive’, ‘tiles’, ‘power lines’, ‘doors’, ‘island’, ‘star’, ‘catcher’, ‘dock’, ‘vegetables’, ‘windshield’, ‘trash bag’, ‘soap dispenser’, ‘shoes’, ‘trash can’, ‘ropes’, ‘wire’, ‘sofa’, ‘sticker’, ‘foot’, ‘arrow’, ‘birds’, ‘hill’, ‘ring’, ‘cabin’, ‘buoy’, ‘graffiti’, ‘stairs’, ‘paint’, ‘antelope’, ‘cabinets’, ‘toothbrush’, ‘walls’, ‘mask’, ‘jars’, ‘lab coat’, ‘plates’, ‘sheep’, ‘hay’, ‘paddle’, ‘toothbrushes’, ‘parking lot’, ‘soup’, ‘dressing’, ‘leaf’, ‘necklace’, ‘baseball bat’, ‘fire truck’, ‘cows’, ‘cucumber’, ‘bucket’, ‘beet’, ‘drawer’, ‘cabbage’, ‘neck’, ‘broccoli’, ‘sauce’, ‘roll’, ‘speakers’, ‘crate’, ‘napkins’, ‘spoon’, ‘controllers’, ‘path’, ‘rope’, ‘bikes’, ‘tires’, ‘highway’, ‘paper towel’, ‘mango’, ‘juice’, ‘rocks’, ‘candles’, ‘entertainment center’, ‘staircase’, ‘dresser’, ‘clothes’, ‘flags’, ‘t-shirt’, ‘wool’, ‘vegetable’, ‘cord’, ‘radiator’, ‘table lamp’, ‘tank top’, ‘onions’, ‘onion’, ‘strawberries’, ‘hose’, ‘wristband’, ‘mane’, ‘boot’, ‘desk lamp’, ‘swimsuit’, ‘chicken’, ‘stick’, ‘tag’, ‘paws’, ‘sock’, ‘twigs’, ‘horns’, ‘train tracks’, ‘pepper’, ‘apron’, ‘lock’, ‘clock tower’, ‘tray’, ‘hills’, ‘life preserver’, ‘traffic lights’, ‘sail’, ‘beef’, ‘macaroni’, ‘pasta’, ‘egg’, ‘beer bottle’, ‘grill’, ‘crown’, ‘dragon’, ‘sailboat’, ‘map’, ‘beans’, ‘potatoes’, ‘sausage’, ‘mouse pad’, ‘coffee cup’, ‘suv’, ‘pear’, ‘gate’, ‘hillside’, ‘headphones’, ‘tissue’, ‘sandal’, ‘bricks’, ‘statue’, ‘stones’, ‘berries’, ‘butter’, ‘pasture’, ‘blinds’, ‘bookshelf’, ‘poster’, ‘ham’, ‘donut’, ‘nightstand’, ‘spray can’, ‘barn’, ‘baseball players’, ‘cds’, ‘horn’, ‘cigarettes’, ‘lemon’, ‘dessert’, ‘tape’, ‘heel’, ‘gravel’, ‘apartment’, ‘hook’, ‘pouch’, ‘pastry’, ‘soap bottle’, ‘walkway’, ‘pig’, ‘tongs’, ‘sheets’, ‘doorway’, ‘carriage’, ‘sculpture’, ‘logs’, ‘pond’, ‘dumpster’, ‘rhino’, ‘icing’, ‘pizza slice’, ‘blueberry’, ‘jar’, ‘pie’, ‘pine tree’, ‘saucer’, ‘tomatoes’, ‘headband’, ‘lawn’, ‘beads’, ‘planter’, ‘lamps’, ‘mannequin’, ‘bleachers’, ‘cutting board’, ‘toaster’, ‘roadway’, ‘pine trees’, ‘envelope’, ‘ski lift’, ‘grapefruit’, ‘motorcycles’, ‘seeds’, ‘antenna’, ‘shower curtain’, ‘cupboard’, ‘toilet lid’, ‘balcony’, ‘chimney’, ‘toaster oven’, ‘straw’, ‘ear buds’, ‘wristwatch’, ‘carts’, ‘comforter’, ‘calculator’, ‘beak’, ‘dome’, ‘keypad’, ‘feeder’, ‘wallpaper’, ‘blender’, ‘mailbox’, ‘pumpkin’, ‘cables’, ‘boys’, ‘light fixture’, ‘eagle’, ‘radio’, ‘lamp shade’, ‘alarm clock’, ‘bar stool’, ‘pancakes’, ‘bandana’, ‘parrot’, ‘sugar packet’, ‘caramel’, ‘croissant’, ‘blouse’, ‘cake stand’, ‘side table’, ‘minivan’, ‘american flag’, ‘seat belt’, ‘wardrobe’, ‘rubber duck’, ‘pipes’, ‘thumb’, ‘tea kettle’, ‘athletic shoe’, ‘newspaper’, ‘heater’, ‘bedspread’, ‘computer desk’, ‘tags’, ‘marker’, ‘broth’, ‘trunks’, ‘fans’, ‘paper container’, ‘earphones’, ‘bikini’, ‘burner’, ‘light bulb’, ‘trailer’, ‘door frame’, ‘hallway’, ‘strawberry’

Objects:
A **knife** is present in the image.
A **rice** is present in the image.
A **countertop** is present in the image.
A **carrots** is present in the image.

Attributes:
The material of the knife present in the image is **stainless steel.**
The material of the knife present in the image is **porcelain**.
The material of the knife present in the image is **plastic**.
The material of the knife present in the image is **ceramic**.

Relations:
The spatial relation between the knife and candle holder is that the knife is **to the right of** the candle holder.
The spatial relation between the knife and candle holder is that the knife is **at** the candle holder.
The spatial relation between the knife and candle holder is that the knife is **between** the candle holder.
The spatial relation between the knife and candle holder is that the knife is **on top of** the candle holder.

Objects:
A **stuffed bear** is present in the image.
A **vehicles** is present in the image.
A **license** plate is present in the image.
A **sign** is present in the image.

Attributes:
The color of the stuffed bear present in the image is **white**.
The color of the stuffed bear present in the image is **striped**.
The color of the stuffed bear present in the image is **pink**.
The color of the stuffed bear present in the image is **purple**.

Relations:
The spatial relation between the stuffed bear and desk is that the stuffed bear is **near** the desk.
The spatial relation between the stuffed bear and desk is that the stuffed bear is **under** the desk.
The spatial relation between the stuffed bear and desk is that the stuffed bear is **on** the desk.
The spatial relation between the stuffed bear and desk is that the stuffed bear is **of** the desk.

Figure C.2. GHOST object-level composition triplet. Green: Positive sentence, Black: Negative sentence

GHOST Unique Attributes

Attribute Names
‘white’, ‘striped’, ‘pink’, ‘purple’, ‘red’, ‘bright’, ‘brown’, ‘green’, ‘gray’, ‘small’, ‘high’, ‘long’, ‘tall’, ‘square’, ‘diamond’, ‘rectangular’, ‘turned’, ‘gold’, ‘yellow’, ‘black and white’, ‘dark’, ‘black’, ‘large’, ‘double decker’, ‘wide’, ‘orange’, ‘colorful’, ‘thin’, ‘low’, ‘bare’, ‘wet’, ‘fresh’, ‘hanging’, ‘blue’, ‘warm’, ‘clear’, ‘happy’, ‘sad’, ‘crying’, ‘silver’, ‘tan’, ‘lit’, ‘used’, ‘painted’, ‘beige’, ‘short’, ‘sunny’, ‘light’, ‘dirty’, ‘covered’, ‘unripe’, ‘walking’, ‘eating’, ‘snowy’, ‘ocean’, ‘deep’, ‘driving’, ‘waiting’, ‘snow’, ‘parked’, ‘military’, ‘paved’, ‘safety’, ‘rounded’, ‘metal’, ‘concrete’, ‘soft’, ‘gravel’, ‘patchy’, ‘wood’, ‘tail’, ‘shiny’, ‘leather’, ‘plastic’, ‘steel’, ‘crispy’, ‘bamboo’, ‘wrinkly’, ‘clay’, ‘knit’, ‘grassy’, ‘public’, ‘sitting’, ‘asphalt’, ‘office’, ‘up’, ‘kitchen’, ‘stone’, ‘brick’, ‘dark brown’, ‘cardboard’, ‘cheese’, ‘tiled’, ‘ceramic’, ‘filled’, ‘clean’, ‘dry’, ‘sliced’, ‘playing’, ‘healthy’, ‘flying’, ‘pointing’, ‘empty’, ‘open’, ‘unpeeled’, ‘peeling’, ‘bald’, ‘baby’, ‘riding’, ‘running’, ‘evergreen’, ‘posing’, ‘cloudy’, ‘upper’, ‘skiing’, ‘wrist’, ‘skating’, ‘connected’, ‘ski’, ‘little’, ‘american’, ‘docked’, ‘urban’, ‘heavy’, ‘bathroom’, ‘cloth’, ‘formal’, ‘busy’, ‘strong’, ‘delicious’, ‘designed’, ‘dark blue’, ‘textured’, ‘mesh’, ‘digital’, ‘smooth’, ‘printed’, ‘off’, ‘lying’, ‘skinny’, ‘narrow’, ‘elderly’, ‘computer’, ‘electric’, ‘triangle’, ‘circle’, ‘octangon’, ‘hard’, ‘stainless steel’, ‘paper’, ‘straw’, ‘glass’, ‘plaid’, ‘shirtless’, ‘tennis’, ‘outdoors’, ‘reading’, ‘toilet’, ‘baseball’, ‘sharp’, ‘jagged’, ‘uneven’, ‘spread’, ‘new’, ‘floral’, ‘broken’, ‘chocolate’, ‘blurry’, ‘asian’, ‘female’, ‘on’, ‘pizza’, ‘aluminum’, ‘halved’, ‘crumbled’, ‘rusty’, ‘piled’, ‘water’, ‘garbage’, ‘wild’, ‘commercial’, ‘curly’, ‘vertical’, ‘round’, ‘porcelain’, ‘worn’, ‘antique’, ‘complete’, ‘swinging’, ‘surfing’, ‘crisp’, ‘dried’, ‘cooking’, ‘loose’, ‘rippling’, ‘rimmed’, ‘tilted’, ‘resting’, ‘closed’, ‘patterned’, ‘attached’, ‘made’, ‘gas’, ‘full’, ‘mounted’, ‘down’, ‘quilted’, ‘still’, ‘rough’, ‘chain-link’, ‘chrome’, ‘brass’, ‘sunlit’, ‘leafy’, ‘street’, ‘wispy’, ‘puffy’, ‘raised’, ‘real’, ‘toy’, ‘tiny’, ‘fluffy’, ‘illuminated’, ‘lined’, ‘standing’, ‘damaged’, ‘leafless’, ‘eaten’, ‘jumping’, ‘shaped’, ‘curved’, ‘flat’, ‘straight’, ‘cut’, ‘male’, ‘short sleeved’, ‘upside down’, ‘skateboarding’, ‘grazing’, ‘old’, ‘fallen’, ‘professional’, ‘grouped’, ‘cooked’, ‘stacked’, ‘kneeling’, ‘drinking’, ‘pointy’, ‘beer’, ‘bent’, ‘curled’, ‘chopped’, ‘melted’, ‘fake’, ‘stuffed’, ‘talking’, ‘coffee’, ‘rubber’, ‘hardwood’, ‘ripe’, ‘grilled’, ‘light brown’, ‘hitting’, ‘indoors’, ‘calm’, ‘staring’, ‘khaki’, ‘crystal’, ‘wavy’, ‘blond’, ‘marble’, ‘rolled’, ‘twisted’, ‘arched’, ‘triangular’, ‘swimming’, ‘young’, ‘squatting’, ‘simple’, ‘artificial’, ‘folded’, ‘floating’, ‘sleeping’, ‘crashing’, ‘wool’, ‘shut’, ‘elevated’, ‘nike’, ‘creamy’, ‘tin’, ‘huge’, ‘sandy’, ‘shallow’, ‘fried’, ‘sprinkled’, ‘wrapped’, ‘bushy’, ‘muddy’, ‘crooked’, ‘fuzzy’, ‘burnt’, ‘apple’, ‘wine’, ‘light colored’, ‘glazed’, ‘wii’, ‘pretty’, ‘funny’, ‘vast’, ‘spiral’, ‘toasted’, ‘baked’, ‘decorative’, ‘ornate’, ‘intricate’, ‘fancy’, ‘messy’, ‘homemade’, ‘chinese’, ‘batting’, ‘opaque’, ‘winter’, ‘smiling’, ‘furry’, ‘plush’, ‘dense’, ‘fat’, ‘hairy’, ‘adult’, ‘crouched’, ‘adidas’, ‘looking up’, ‘roman’, ‘lush’, ‘granite’, ‘tied’, ‘curious’, ‘christmas’, ‘angled’, ‘spinning’, ‘dark colored’, ‘bronze’, ‘neon’, ‘rocky’, ‘overhead’, ‘pine’, ‘decorated’, ‘sweet’, ‘sleeveless’, ‘wicker’, ‘suspended’, ‘beautiful’, ‘dusty’, ‘stained’, ‘thick’, ‘plain’, ‘breakable’, ‘rainbow colored’, ‘dotted’, ‘frosted’, ‘reflective’, ‘barefoot’, ‘dull’, ‘shining’, ‘rustic’, ‘lighted’, ‘sparse’, ‘blooming’, ‘ivory’, ‘weathered’, ‘looking down’, ‘transparent’, ‘murky’, ‘clumped’, ‘copper’, ‘sliding’, ‘outstretched’, ‘snowboarding’, ‘cracked’, ‘performing trick’, ‘framed’, ‘fenced’, ‘shaded’, ‘crouching’, ‘woven’, ‘cream colored’, ‘chipped’, ‘carved’, ‘bending’, ‘checkered’, ‘capital’, ‘crossed’, ‘edged’, ‘misty’, ‘octagonal’, ‘shredded’, ‘uncooked’, ‘long sleeved’, ‘denim’, ‘trimmed’, ‘maroon’, ‘forested’

GHOST Unique Relations

Relation Names
‘near’, ‘under’, ‘on’, ‘of’, ‘to the right of’, ‘in’, ‘in front of’, ‘to the left of’, ‘at’, ‘behind’, ‘around’, ‘above’, ‘below’, ‘between’, ‘on top of’, ‘surrounded by’, ‘close to’, ‘worn by’, ‘on the front of’, ‘on the side of’, ‘hanging over’, ‘inside’, ‘standing in’, ‘next to’, ‘by’, ‘besides’, ‘wearing’, ‘flying’, ‘watching’, ‘looking at’, ‘covered in’, ‘leaning on’, ‘holding’, ‘eating’, ‘carrying’, ‘with’, ‘playing’, ‘observing’, ‘working on’, ‘sitting on’, ‘riding’, ‘cutting’, ‘making’, ‘standing behind’, ‘swinging’, ‘riding on’, ‘sitting in’, ‘lying on’, ‘covered by’, ‘on the other side of’, ‘throwing’, ‘covering’, ‘using’, ‘reading’, ‘perched on’, ‘standing on’, ‘mounted on’, ‘driving’, ‘pulling’, ‘sitting behind’, ‘sitting beside’, ‘hitting’, ‘sitting on top of’, ‘on the back of’, ‘standing in front of’, ‘hanging from’, ‘sleeping on’, ‘beside’, ‘underneath’, ‘playing with’, ‘walking in’, ‘crossing’, ‘holding onto’, ‘hanging off’, ‘attached to’, ‘touching’, ‘contain’, ‘walking on’, ‘talking on’, ‘pulled by’, ‘traveling on’, ‘reflecting in’, ‘going down

D. GHOST Dataset Visualization

A compositional triplet consists of an object type, an attribute, and a relation. Each component of the compositional triplet includes one positive statement and three negative statements. From GHOST, we visualize two images with a compositional triplet for each image in Figure C.2.

E. MLLM Evaluation Visualization

We show various MLLM predictions to our GHOST examples in the following Tables.

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"stuffed bear"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
stuffed bear	True	True	True	True	True	True	True	True	True
license plate	False	False	False	False	False	False	False	False	False
sign	False	False	True	False	True	False	False	False	False
vehicles	False	False	False	False	False	True	False	False	False
Attributes									
white	True	True	True	True	True	True	True	True	True
pink	False	False	False	False	True	False	False	False	False
purple	False	False	False	False	False	False	False	False	False
striped	False	False	False	False	False	False	False	False	False
Relations									
near	True	False	True	True	True	True	True	True	True
of	False	False	True	False	True	False	False	True	False
on	False	False	False	False	True	True	False	True	False
under	False	False	False	False	True	True	True	False	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"monitor"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
monitor	True	True	True	True	True	True	True	True	True
meal	False	False	False	False	False	False	False	False	False
mountain	False	False	False	False	False	False	False	False	False
plant	False	False	False	False	False	False	False	False	False
Attributes									
black	True	True	True	True	True	True	False	True	True
black and white	False	False	False	True	False	True	False	False	True
light	False	False	False	True	True	True	False	False	False
silver	False	False	False	False	False	True	False	False	False
Relations									
to the left of	True	True	False	True	False	True	True	True	True
around	False	False	True	True	True	False	False	False	True
at	False	False	True	True	True	False	False	False	False
between	False	False	False	True	True	True	False	False	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"tray"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
tray	True	True	True	True	True	True	True	True	True
book	False	False	False	False	False	False	False	False	False
lamps	False	False	False	False	False	False	False	False	False
shelf	False	False	False	False	True	False	False	False	False
Attributes									
silver	True	True	True	True	True	True	True	True	True
black	False	False	False	False	True	False	False	False	False
brown	False	False	False	True	True	False	False	False	False
striped	False	False	False	False	False	False	False	False	False
Relations									
above	True	True	True	True	True	True	True	False	True
around	False	False	False	True	True	True	True	False	True
inside	False	False	False	True	False	True	True	False	False
near	False	False	True	True	True	True	True	True	True

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"mug"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
mug	True	True	True	True	False	True	True	True	False
books	False	True	False	True	True	True	False	False	True
steps	False	False	False	False	True	False	False	False	False
umbrella	False	False	False	False	False	False	False	False	False
Attributes									
colorful	True	False	False	True	False	True	False	True	True
clear	False	False	False	True	False	False	False	False	False
dark	False	False	True	True	False	True	False	False	False
green	False	False	False	False	False	False	False	False	False
Relations									
to the left of	True	True	True	True	False	True	True	True	False
above	False	True	True	True	True	True	False	False	True
in	False	False	False	False	False	False	False	False	False
in front of	False	True	True	True	True	True	True	True	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"animal"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
animal	True	True	True	True	True	True	True	True	True
camera	False	False	False	True	True	True	False	False	False
paint	False	False	False	False	True	True	False	True	False
snow	False	False	False	False	False	False	False	False	False
Attributes									
black	True	True	True	True	False	True	False	False	False
orange	False	False	False	False	False	False	False	False	False
striped	False	False	False	False	True	True	False	False	False
tan	False	False	False	False	True	False	False	False	False
Relations									
eating	True	True	True	True	True	True	True	True	True
covering	False	False	False	False	True	False	False	False	False
making	False	False	False	False	True	False	False	False	False
sitting on	False	False	True	False	True	False	False	True	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"plate"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
plate	True	True	True	False	True	False	False	True	True
catcher	False	False	False	False	False	False	False	False	False
field	False	False	False	False	False	False	False	False	False
frame	False	False	False	False	True	False	False	False	False
Attributes									
decorative	True	True	True	True	False	True	True	True	True
formal	False	False	False	True	False	True	False	True	False
rustic	False	False	True	True	False	True	False	False	True
simple	False	True	True	True	False	True	True	True	True
Relations									
to the right of	True	True	True	True	True	True	True	True	True
above	False	False	True	True	True	True	True	True	True
at	False	False	False	False	True	False	False	False	False
to the left of	False	False	False	True	True	True	True	True	True

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"man"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
man	True	True	True	True	True	True	True	True	True
giraffes	False	False	False	False	False	False	False	False	False
snow	False	False	False	False	False	False	False	False	False
spots	False	False	True	False	True	True	True	True	False
Attributes									
upside down	True	True	True	True	True	True	True	True	True
cooked	False	False	False	False	False	False	False	False	False
ripe	False	False	False	False	False	False	False	False	False
sleeping	False	False	False	False	False	False	False	False	False
Relations									
with	True	True	True	True	True	True	True	True	True
observing	False	False	False	False	True	False	False	False	False
playing	False	False	False	True	True	True	False	True	False
using	False	False	False	False	True	False	False	False	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"cell phone"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
cell phone	True	True	True	True	True	True	True	True	True
beverage	False	False	True	True	True	True	True	True	True
post	False	False	False	False	True	False	False	False	False
steps	False	False	False	False	True	False	False	False	False
Attributes									
black	True	True	True	True	True	True	True	True	True
blue	False	False	False	False	False	False	False	False	False
bright	False	False	False	True	True	True	False	False	False
warm	False	True	True	False	False	True	False	True	False
Relations									
to the right of	True	True	True	True	True	True	True	True	True
of	False	False	False	False	False	False	False	False	False
on top of	False	False	False	False	False	False	False	False	False
to the left of	False	False	False	True	False	True	False	False	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"shorts"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
shorts	True	True	True	True	True	True	True	True	True
gloves	False	False	False	False	True	False	False	False	False
umbrella	False	False	False	False	False	False	False	False	False
wine	False	False	False	False	False	False	False	False	False
Attributes									
black	True	True	True	True	True	True	False	True	True
black and white	False	False	True	False	True	False	False	False	True
striped	False	False	False	False	False	False	False	False	False
yellow	False	False	False	False	True	False	False	False	False
Relations									
to the left of	True	True	True	True	True	True	True	True	True
behind	False	False	True	False	True	True	True	True	False
between	False	False	True	True	True	False	True	True	False
on	False	False	False	False	False	False	False	False	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"newspaper"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
newspaper	True	True	True	True	True	True	True	True	True
desk	False	False	False	False	False	False	False	False	False
plants	False	False	False	False	False	False	False	False	False
watch	False	True	False	True	True	False	False	False	False
Attributes									
white	True	True	True	True	True	False	False	True	False
colorful	False	False	False	False	True	False	False	False	False
red	False	False	False	False	False	False	False	False	False
striped	False	False	False	False	False	False	False	False	False
Relations									
to the left of	True	True	True	True	True	True	True	True	False
at	False	False	False	False	True	False	False	False	True
behind	False	False	False	False	True	True	True	False	False
to the right of	False	False	False	False	True	True	True	False	False

GHOST Prediction for different MLLM

Task: Hallucination estimation through consistency check for the objects, attributes, and relations of **"headband"** in the image for various MLLM models.
Image:

Predictions: Comparison of ground truth (**GT**) with predictions from different MLLMs.

Feature	GT	GPT-4o	GeminiPro1-5	Phi-3.5-V	Idefics-9B	LLaVA-1.5-13B	VILA-1.5-13B	LLaVA-OV-7B	MiniCPM-LLaMA3
Objects									
headband	True	True	True	True	True	True	True	True	True
counter	False	False	False	False	True	False	False	False	False
fruit	False	False	False	False	False	False	False	False	False
giraffes	False	False	False	False	False	False	False	False	False
Attributes									
brown	True	True	True	True	True	True	True	True	True
black	False	False	False	False	True	True	False	False	False
blue	False	False	False	False	True	False	False	False	False
green	False	False	False	False	False	False	False	False	False
Relations									
to the right of	True	False	True	True	True	True	True	True	False
behind	False	False	True	False	False	True	True	False	False
on	False	False	False	False	False	False	False	False	False
to the left of	False	False	True	True	False	True	True	True	False

References

[1] Gpt-4: Gpt-4o ("o" for "omni"). https://openai.com/index/hello-gpt-4o/. com/news/claude-3-haiku. 4

[2] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2, 4

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2

[4] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*, 2024. 2

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024. 2

[6] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024. 1

[7] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 2

[8] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 2, 4

[9] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[10] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. arXiv2405.01483. 2

[11] 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. corr abs/1602.07332, 2016. 5

[12] Hugo Laurencon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, et al. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023. 2

[13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 4

[14] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023. 2, 4

[15] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 4

[16] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv:2408.15998*, 2024. 2

[17] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2024. 2

[18] 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 preprint arXiv:2408.01800*, 2024. 2