What's in Common? Multimodal Models Hallucinate When Reasoning Across Scenes

Candace Ross **Florian Bordes**

Adina Williams **Polina Kirichenko** **Mark Ibrahim**

FAIR at Meta

{ccross, fbordes, adinawilliams, polkirichenko, marksibrahim}@meta.com

Abstract

Multimodal language models possess a remarkable ability to handle an open-vocabulary worth of objects. Yet the best models still suffer from hallucinations when reasoning about scenes in the real world, revealing a gap between their seemingly strong performance on existing perception benchmarks that are saturating and their reasoning in the real world. To address this gap, we build a novel benchmark of in-the-wild scenes that we call `Common-O Bench`. With more than 10.5k examples using exclusively new images not found in web training data to avoid contamination, `Common-O Bench` goes beyond just perception, inspired by cognitive tests for humans, to probe reasoning across scenes by asking "what's in common?". We evaluate leading multimodal language models, including models specifically trained to reason. We find that perceiving objects in single images is easy for most models, yet reasoning across scenes is very challenging even for the best models, including reasoning models. Despite saturating many leaderboards focusing on perception, the best performing model only achieves 35% on `Common-O Bench`—and on `Common-O Complex`, consisting of more complex scenes, the best model achieves only 1%. Curiously, we find models are more prone to hallucinate when similar objects are present in the scene, suggesting models may be relying on object co-occurrence seen during training. Among the models we evaluated, we found scale can provide modest improvements while models explicitly trained with multi-image inputs show bigger improvements, suggesting scaled multi-image training may offer promise. We make our benchmark publicly available to spur research into the challenge of hallucination when reasoning across scenes at `https://huggingface.co/datasets/facebook/Common-O`.

1 Introduction

Multimodal models today have begun saturating visual perception leaderboards. For example, on benchmarks such as CLEVR (Johnson et al., 2016), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), MMBench (Liu et al., 2023), and Seed-Bench (Li et al., 2024b), top-performing models achieve an accuracy of 80%-90% (Zhang et al., 2024). However, despite these impressive results, there is a growing concern that these benchmarks may not accurately reflect the performance of models in real-world scenarios. In fact, research has shown that models often struggle to generalize to new, unseen data, and are prone to hallucinating objects that are not present in the scene (Guan et al., 2023).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

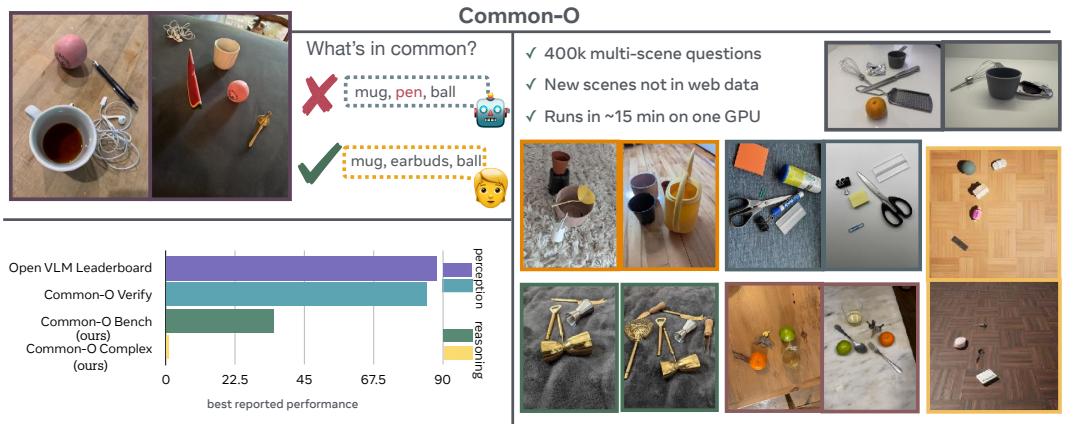


Figure 1: **Reasoning across scenes is an open challenge for today's best multimodal models.** We show the best performance from the Open VLM leaderboard on MMBench and single image evaluations from our benchmark illustrating saturation for perception tasks.

One of the main reasons for this gap between benchmark performance and real-world performance is the contamination between benchmarks and training data (Chen et al., 2024). Many benchmarks are constructed from web data, which is also used to train models. This means that models are often trained on data that is similar to the benchmark data, leading to an overestimate of their performance. Another factor contributing to the gap is the lack of benchmarks that capture the complexity and variability of real-world scenes. Many benchmarks use simplified geometric visuals or static scenes; while useful for static tasks such as chart understanding, these do not reflect the dynamic and evolving nature of real-world scenes.

The ability to reason across complex scenes containing multiple objects is a fundamental aspect of human cognition. Research in cognitive science has shown object identification in scenes is a key component of cognitive function—and that deficits in this ability are a hallmark of cognitive decline, such as in Alzheimer's disease (Takechi & Dodge, 2010). Furthermore, studies have demonstrated that the brain's ability to understand the relationships between objects in a scene is closely tied to its ability to understand the scene as a whole (Frankland & Greene, 2015). These findings suggest that a benchmark that requires models to reason about complex, dynamic scenes is crucial step towards deploying reliable models in the real world.

To address these challenges, we introduce a new multi-image benchmark,`Common-O Bench`, designed to test models' ability to reason *across dynamic scenes* in a way more similar to human reasoning. Our benchmark includes scenes containing multiple objects with varying lighting conditions, and complex backgrounds that requires models to reason about the relationships between objects across distinct scenes. We choose scenes with up to 7 objects as our default setting, inspired by the 1956 classic, putative heuristic constraint on human memory, colloquially described as the "Magical Number Seven, Plus or Minus Two" (Miller 1956; Baddeley 1994; Cowan et al. 2007, i.a.). `Common-O Bench` comprises both real and synthetic data, allowing for more flexibility in our evaluation, as we can sample a wide range of object-background combinations that are typical in real data. We also provide a non-overlapping fully synthetic challenge set, `Common-O Complex` that spans up to 16 objects per scene, increasing scene complexity appreciably. In both `Common-O Bench` and `Common-O Complex`, we intentionally provide multiple camera points of view of a given scene, reflecting the diversity found in the real world.

We find that despite being able to recognize objects in individual scenes, state-of-the-art models struggle to reason across scenes. The best performing model, GPT-4o, achieves only 35% on `Common-O Bench`, highlighting reasoning across scenes as open challenge, in stark contrast to the saturation observed for other multimodal benchmarks. For the more challenging set `Common-O Complex`, the best performing model achieve $<$1%. Curiously, we find models hallucination is *pervasive*, with at least one 1 object hallucinated 53% of the time and 2+ as often as 23% of the time.

Our findings have important implications for the development of multimodal models. We find models trained with multi-image inputs achieve higher performance and scale can yield marginal benefits, yet even the best multi-image large scale models struggle highlighting the need for models to be designed with real-world scenes in mind. This requires a fundamental shift in the way models are designed and trained, and underscores the need for more research in this area. We release `Common-O Bench` and `Common-O Complex` to mark a new challenge in multimodal models' ability to reason across scenes that we hope could unlock new frontiers in real world applications [1].

2 Related Work

Many works have aimed to evaluate model performance on visual reasoning. We summarize our contributions relative to existing benchmarks in terms of multi-image capability, scale, and saturation in Table 1. Our dataset is larger in size, captures multi-image reasoning across scenes inspired by human cognitive tests, and stands out in terms of not relying on existing web datasets, thereby avoiding possible training data contamination or object resemblance. Together, these factors make our benchmark much more challenging relative to existing benchmarks where performance has saturated.

	Benchmark	Multi-image	Multi-scene	Size	Source	SOTA
Abstract	NTSEBENCH	×	×	2.7k	Web	88.9%
	MathVista	×	×	6k	Existing & new	80.9%
	MMIU Objective Semantic	✓	✓	1.2k	Existing	55.7%
	ReMI	✓	✓	2.6k	Synthetic	50.5%
Hallu.	POPE	×	×	9k	Existing	91.0%
	HallusionBench	✓	✓	591	Synthetic & cartoon	67.1%
Real	MMBench	×	×	1784	Web	88.3%
	NLVR2	✓	✓	13.9k	Web	80.3%
	GQA	✓	✓	3.4k	Existing	74.6%
	SEED-Bench-2	✓	✓	660	Existing	73.1%
	MUIRBench	✓	✓	536	Existing & new	68%
Ours	`Common-O Bench`	✓	✓	10k	New	35%
	`Common-O Complex`	✓	✓	12k	New	1%

Table 1: Existing benchmark datasets targeting abstract reasoning, hallucination ('Hallu.'), and real image reasoning) are insufficient due to saturation, and/or failure to target multi-image and/or multi-scene reasoning. Existing datasets targeting multi-image and multi-scene reasoning exist but have saturated (NLVR2, GQA). Those that have not saturated are relatively small (SEED-Bench-2, MUIRBench, HallusionBench, ReMI). Abstract benchmarks mostly focus on abstract geometric reasoning in puzzles/charts rather than real scenes or extract frames from videos.

Perception. Many benchmarks include composite measures that focus on single object-centric perception such as object classification (Deng et al., 2009; Lin et al., 2014) and attributes or relations of objects (Al-Tahan et al., 2024; Dumpala et al., 2024). As part of perception, researchers have also focused on the issue of hallucination where models describe objects that are not present in scenes (Li et al., 2023b; Guan et al., 2023). Instruction following (Li et al., 2023a) for perception tasks using single images is another area where diversity, quality, and creativity of answers is important. Recent efforts to benchmark multimodal model have relied on larger composite suites of benchmarks that span several tasks such as recognition, OCR, counting, visual question answering, and object attributes etc. (Yu et al., 2023; Liu et al., 2023; Li et al., 2024b).

Abstract reasoning in charts, geometric sketches, and puzzles. Relative to the improved performance on real world perception tasks, multimodal models exhibit degraded performance on abstract visual puzzles that involve straight-forward reasoning. For example, Rahmanzadehgervi et al. (2025) show multimodal models lag considerably behind humans at identifying simple tasks such as whether two circles overlap with Huang et al. (2025) showing similar conclusions on visual arithmetic. Similarly, Wüst et al. (2025); Jiang et al.; Ullman (2024); Kraaijveld et al. (2024) probe whether models

[1] Datasets are available at `https://huggingface.co/datasets/facebook/Common-O`

can solve basic visual logical puzzles that involve outlines of geometric shapes, illusions, and lateral thinking. Pandya et al. (2025) construct a dataset of 2.7k multiple choice questions from the national exam in India that involve geometry and visual reasoning questions from graphs. Hemmat et al. evaluates whether multimodal models can perceive abstract shapes, a key aspect of human visual perception. Sampat et al. (2024) assess whether multimodal models can solve NLP and visual tasks jointly. Lin et al. (2024) studies comparisons across pairs of synthetically generated CAD images. Most similar to our work is the objective high-level semantic task from MMIU, which consist of 1.1k examples from existing datasets focused on semantic correspondence such as BLINK (Fu et al., 2024) and MISC210K (Sun et al., 2023), spotting the difference (Jhamtani & Berg-Kirkpatrick, 2018) or abstract puzzles from datasets such as NLVR2 (Suhr et al., 2019). We build on this setup to focus on reasoning about object commonality across scenes at larger scales.

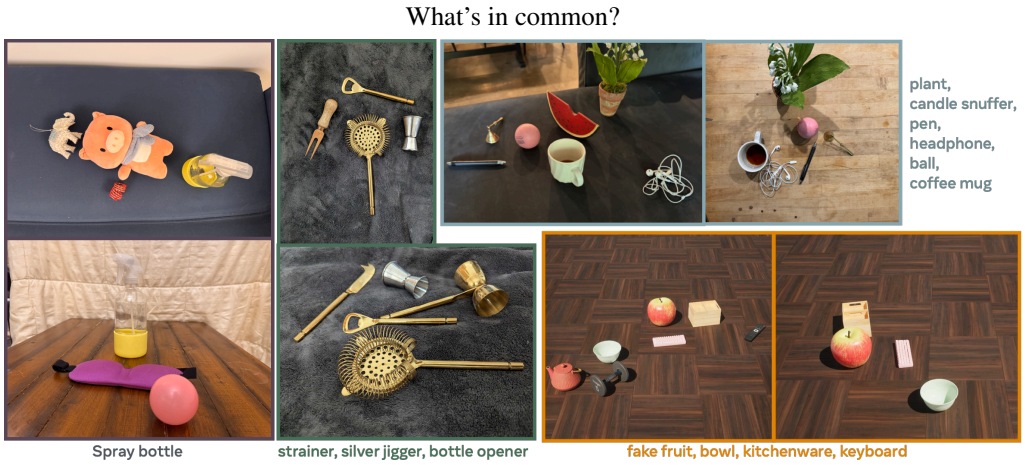
Measuring reasoning using single image benchmarks. The prior benchmarks reveal abstract reasoning may be a challenge for multimodal models hinting at a possible reason for the observed gap in real world performance multimodal models. Many works attempted to measure the gap between real world capabilities and benchmark performance by focusing on robustness (Geirhos et al., 2022; Gabbay et al., 2021; Hendrycks et al., 2021). For example, Richards et al. (2023) measures the in the wild robustness gap for household object classification across geographies. Another approach to capture the real world versus standard benchmark gap is to explicitly mine or generate challenging images (Tong et al., 2024a,b; Wang et al., 2025). A first step to reasoning beyond perception is compositionality. Several works have studied whether multimodal models can understand and compose attributes and objects (Johnson et al., 2016; Thrush et al., 2022; Yuksekgonul et al., 2022; Krojer et al., 2022; Wu et al., 2025). Some have even explored single-image reasoning in adversarial settings (Li et al., 2021; Sheng et al., 2021) and memes (Kiela et al., 2020; Suryawanshi & Chakravarthi, 2021). Yet, real world scene understanding requires reasoning beyond basic single-image settings.

Multi-image reasoning across natural scenes. To go beyond perception, *generalization in the real world requires reasoning across scenes*. Aggregate benchmark such as VHELM contain multi-image reasoning tasks (Lee et al., 2024b), many of which are derived from geometry style puzzles akin to those described above. The real world image reasoning task in VHELM is based on GQA, which is a dataset constructed from objects in the popular Visual Genome dataset available in web training data (Agrawal et al., 2022). The authors use 11k images from the GQA validation set in their evaluations. There are also multi-image binary tasks with image selection (Hu et al., 2019), predicting whether captions are true of images (Suhr et al., 2017, 2019), and visual haystacks (Wu et al., 2024a) that focuses on retrieval as well as visual question answering based on a large number of images (up to 10k). Other benchmarks Meng et al. (2024); Fu et al. (2024) also focus on multi-image tasks, showing even models that excel at single image tasks struggle on multi-image tasks such as visual correspondence, semantic correspondence, and multi-view reasoning of the same scene across multiple images. Other tasks include visual similarity, relative depth, and functional correspondence in the same image. However, as shown in Table 1, these multi-image real scene benchmarks rely on mining image from the web or existing datasets, which both limits their size and introduces possible training data contamination. We observe the best reported performance even on multi-image benchmarks is quite high 68-88.3%.

3 Methods

3.1 Dataset Construction

`Common-O Bench` is designed to test the ability of models to reason about complex, dynamic scenes in a way that is similar to human reasoning. `Common-O Bench` consists of 10.5k examples, representing different scenes containing 3 or more objects with diverse background and viewpoints. configurations of objects. Every example in the dataset consists two images, which can be either real (45%) or synthetic (55%). The real images, we took images ourselves, to ensure that each image is completely new and unique, with no issues of contamination in web or existing data used for training. Real images were taken by four experts in machine learning with no particular photography training. Image-takers followed a fairly simple data creation procedure where images were grouped in sets and placed arbitrarily against backgrounds to generate test data. We do not include any images of people



(a) `Common-O Bench`: 10k examples of real and synthetic images, with scene complexity from 3 to 7 objects.

(b) `Common-O Complex`: 12k containing synthetic images only, ranging in complexity from 8 to 16 objects.

Figure 2: `Common-O Bench` contains real and synthetic images of objects in different orientations and configurations. These are randomly selected examples from the dataset along with the human ground truth labels for the common object(s) between them.

or proprietary content such as logos. See Appendix A for more details on our image-taking guidelines. The synthetic images were generated using Unreal Engine 5.4 with assets from the Aria Digital Twin Catalog(Dong et al., 2025). We place the objects randomly in the scene and take pictures from different angles. To avoid overlapping between objects we rescale each of them to a given maximum size while keeping their aspect ratio (more detail can be found in Appendix C). We also construct `Common-O Complex` consisting of 12k examples of more complex scenes, containing $8-16$ objects, and being wholly synthetically created using the same video game engine. This allows us to evaluate the ability of models to reason about scenes with varying levels of complexity. We have 129 different objects in `Common-O Bench` and `Common-O Complex`. Using Segment Anything (Kirillov et al., 2023), we find the object in the images ranges from 2-22% of the overall image size. Following Gebru et al. (2021), we include a full dataset card in Appendix E. See Figure 2 for examples from `Common-O Bench` and `Common-O Complex`.

3.2 Evaluation

Task Definition. An input example is defined as $(I_0, I_1, \mathcal{O}_{\text{choices}}, \mathcal{O}_{\text{in\_common}})$ where:

- I_0, I_1 are the two images I_0, I_1
- $\mathcal{O}_{\text{choices}}$ is a set of candidate objects $\mathcal{O}_{\text{choices}}$
- $\mathcal{O}_{\text{in\_common}}$ is the set of ground truth objects in common between the images.

Models are tasked with predicting the common objects $\mathcal{O}_{\text{in\_common}}$. We format the data $(I_0, I_1, \mathcal{O}_{\text{choices}})$ into model input.

To isolate perception from reasoning capabilities, we conduct a single-image evaluation as well. Models receive one image and a binary question (*"Is <object> in this image?"*), testing basic object recognition. Strong performance here suggests failures in multi-image setups stem from reasoning limitations rather than perception deficits. This controlled comparison enables clearer analysis of cross-image reasoning abilities. We also performed human annotations with 4 expert annotators who are authors using 100 randomly sampled examples (each reviewed by at least two annotators). We reach 84% annotation agreement.

Metrics. We assess performance through two complementary metrics. First, **accuracy** measures strict correctness, requiring an exact match between predicted ($\mathcal{O}_{\text{pred}}$) and ground truth ($\mathcal{O}_{\text{common}}$) object sets. Second, **hallucination rate** quantifies how often model respond with an object that is not present. Speficially, hallucination measures the false positive predictions, calculated as the ratio of incorrectly predicted objects to total choices: $\frac{|\mathcal{O}_{\text{pred}} \setminus \mathcal{O}_{\text{common}}|}{|\mathcal{O}_{\text{choices}}|}$. This combination enables evaluation of both precision and recall in model predictions.

Models. We benchmark a diverse array of multimodal models spanning different architectural families and scales. Openly available models include LLaVA-OneVision (7B, 72B) (Li et al., 2024a), DeepSeek-VL2 (Small/Base) (Wu et al., 2024b), LLaMA-V-o1 (Thawakar et al., 2025), Qwen2.5-VL (Bai et al., 2025), LLaMA-4 Scout Instruct (Meta), PerceptionLM (3B/8B) (Cho et al., 2025) and QVQ-72B-Preview (Team, 2024). The closed-source GPT-4o is also evaluated[2]. Our implementation uses HuggingFace Transformers (Wolf et al., 2020) for LLaMA-V-o1, the Perception Models GitHub repository[3] for PerceptionLM, and vLLM (Kwon et al., 2023) for remaining models. We ran all models locally, on single node with 8 A100s GPUs, except for GPT-4o, which is only available through the API. All use greedy decoding with default parameters (temperature=1, top-p=1) unless specified otherwise. Images are resized, maintaining the aspect ratio, with the smallest size of 384px. For models not explicitly trained for multi-image input—Llama 3.2, LlamaV-o1, PerceptionLM—we first concatenate the two images before passing them to the model as input. [4]

Model Input. The object choices are alphabetized (A, B, C. . .) to leverage models' preference for letter-based responses over other input formats (Long et al., 2024). Outputs must conclude with a comma-separated prediction list, allowing flexible generation formats, including chain-of-thought reasoning (Wei et al., 2022). For models trained for multi-image input, text prompt is:

```
Which objects are present in both images? Select all choices
that are true: {}. You can think of your answer in any way (e.g.
step-by-step) but for the last line of your response, respond only
in this format 'Answer: <letter 1> <letter 2> <letter 3>', e.g.
'Answer: A, B, C'.
```

For models where we first concatenate the input images, the text prompt is:

```
There are two images provided, one on the left and the other on
the right. Which objects are present in both images? Select all
choices that are true: {}. You can think of your answer in any way
(e.g. step-by-step) but for the last line of your response, respond
only in this format 'Answer: <letter 1> <letter 2> <letter 3>',
e.g. 'Answer: A, B, C'.
```

We also tested two additional input prompt formulations, shown in Appendix B. We did not observe a significant performance difference across prompts.

[2]Note that we use a slightly different prompt setup for GPT-4o, where the model predicts object values instead of letters. We provide the full comparison in the appendix.

[3]`https://github.com/facebookresearch/perception_models`

[4]For the best performing open source model, we additionally tested different temperatures and did not observe a significant performance difference. Results are shown in Appendix B

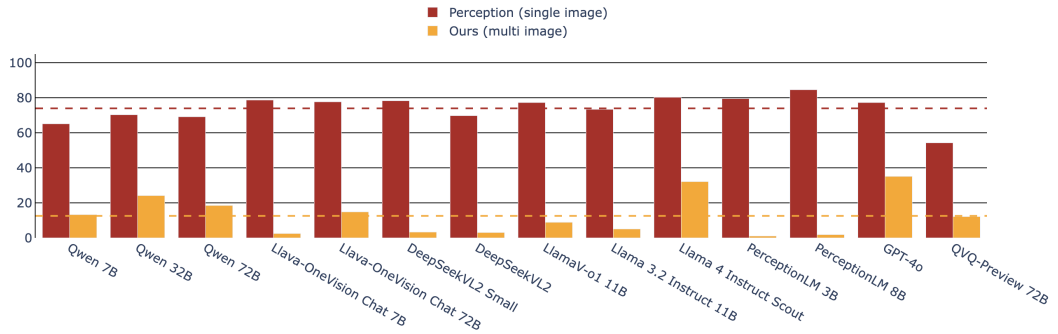
4 Results

4.1 Multimodal models can perceive, but struggle to discern what's in common across scenes.

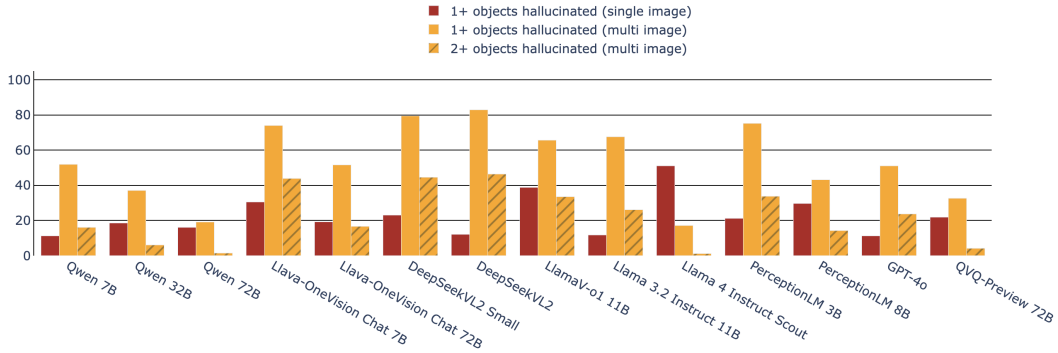
To evaluate the performance of various state-of-the-art models on `Common-O Bench`, we first validate the difficulty of perception using a single image scene setup as shown in Figure 3a. We find all models exhibit strong performance on single-image perception, yet struggle to reason across the same scenes in `Common-O Bench`. The best performing model, GPT-4o, achieves only 35% accuracy with reasoning models performing even worse, highlighting the challenge of reasoning across scenes. Underlying many incorrect answer is a tendency for models to hallucinate objects. We also report standard errors, which we find to $< 0.02\%$, for all models across both single image perception and multi-image reasoning in Appendix B.

4.2 Models often hallucinate objects when reasoning across scenes.

Models are also very likely to hallucinate objects that are not present, which impacts their multi-scene reasoning abilities. One perspective could be that, because of models' *yes*-bias, they're likely to hallucinate objects that are not present when asked about them directly (Zhang et al., 2016; Agrawal et al., 2018; Ross et al., 2024). While models do hallucinate in this single image setting, Figure 3b shows that hallucinate rates are significantly higher during the multi-image reasoning setting (excluding Llama 4, which does not follow this trend). We show an example hallucination in Figure 4. Similarly, hallucinating multiple objects rarely occurs in single image perception, but occurs more often when reasoning across scenes.



(a) *Accuracy (↑):* Models can perform well in perceiving objects in the single image setting, but struggle to reason across scenes in our multi-image setting. The dashed lines show performance averaged across models.



(b) *Rates of hallucination (↓):* Models often hallucinate objects that are not present in the image. Models by and large hallucinate objects significantly more often in the multi-image case; Llama 4 was the only exception. Single image multiple object hallucination rates are not plotted as they were negligible for all models.

Figure 3: Performance for single image object perception **in yellow** and multi-image reasoning **in red** for (a) accuracy and (b) hallucination rates. We observe models with higher accuracy tend to also have lower rates of hallucination. We include a table of these results, along with statistical analyses, in Appendix D.

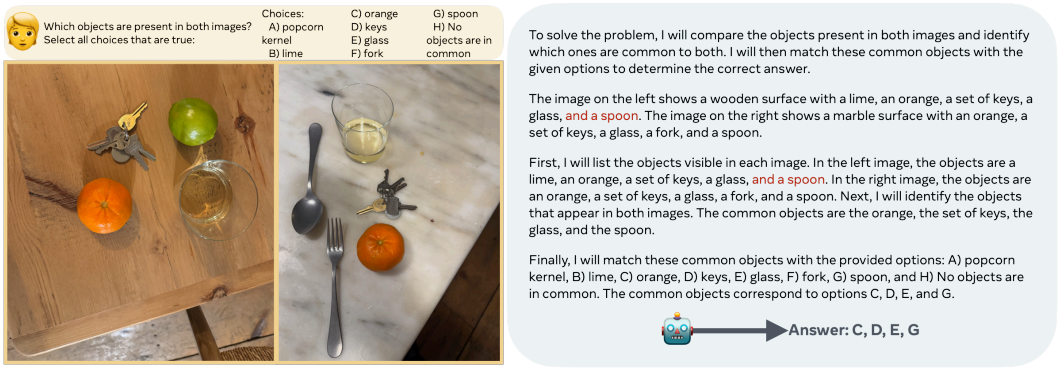


Figure 4: These are two examples of model failures, with the specific failures shown **in red**.

Model performance *severely* degrades for more complex scenes. We use the challenge set, `Common-O Complex`, to explore how models perform with more complex scenes. `Common-O Complex` follows the same design described in Section 3.1, we just extended the scene complexity to N=8 to N=16 objects (see examples in Figure 2b). Across every model, performance severely drops. None of the models that we evaluated gets above $\sim$1% accuracy. We also see very high hallucination rates, with 76% for 1+ objects and 55% for 2+ objects on average across all models. This bolsters our decision to use a default setting of 7 objects as a good primary focus for models.

	Qwen			Llava Chat		DeepSeek		Llama			PLM		QVQ
	7B	32B	72B	7B	72B	Small	Base	V-o1	3.2	4	3B	8B	72B
Acc. (%)	0.1	0	0.01	0.05	0	0.07	0.04	0.1	0.1	0	0	0	0.03

PLM = PerceptionLM

Table 2: On `Common-O Complex`, with the complexity ranging from 8 to 16 objects per scene, model performance *severely* degrades. The best performing models reach <1% accuracy.

When objects are similar, it's harder for models. Next, we test the effects of the *similarity* of the common objects within a set. If objects in images are similar, it may pose a unique challenge for models. For a given set of common objects, $\mathcal{O}_{common}$, we compute an embedding for each object in the set and take the maximum pairwise similarity as a proxy for object similarity. We use the NV-Embed2 embedding model (Lee et al., 2024a), as it was optimized for embedding similarity. We observe that accuracy generally *decreases* as object similarity *increases*, meaning similarity among objects perhaps makes the task of reasoning about commonality more challenging. We validate this statistically by computing the Pearson correlation between similarity of common objects and accuracy, and find 10 or our 13 tested models have statistically significant, negative correlations of small effect size with $|r| >= 0.3$ (see Appendix, Table 6 for full results).

4.3 How do real and synthetic images compare?

We compare model performance on the real images versus synthetic images. To do this, we focus on `Common-O Bench` results, and subset the dataset according to whether the examples were real or synthetic. We find that synthetic images are generally more challenging for models than real images, with less of a gap between the performances on the two data subtypes for models that were less performant overall on `Common-O Bench` (see Figure 5 for full results). Though the synthetic images are similar to real images in several respects, having the same scene complexity and using multiple camera orientations per configuration, the synthetic images have the potential to be more diverse in backgrounds and object sizes. This increased difficulty may be due to a domain shift from models' training data. We used diverse backgrounds (*e.g.* green marble, concrete, aluminum) and relative object sizes that are less common in the real world (*e.g.* a rubber duck being the same size as a remote). Additionally, because data contamination is difficult to avoid once benchmarks are

openly available on the internet, our results show the benefit of leveraging synthetic data without compromising on image difficulty or quality.

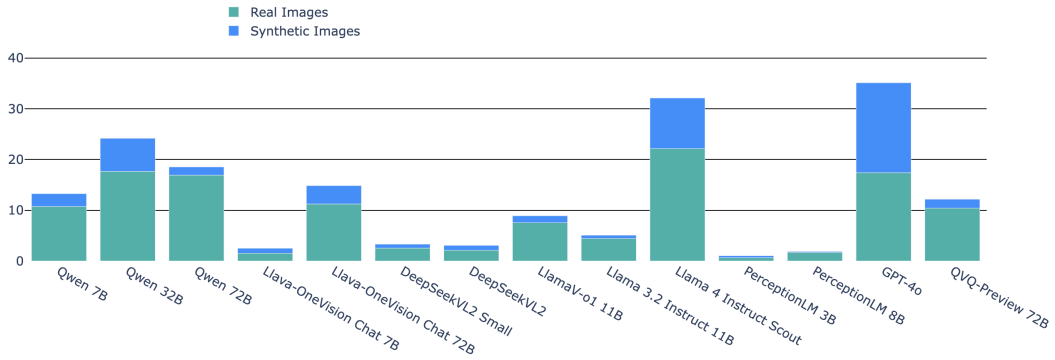


Figure 5: Performance on `Common-O Bench` subsetted according to whether example image pairs are real or synthetic. The height of each bar represents the total accuracy on `Common-O Bench`: the **green area** of the bar represents the contribution of the real image accuracy, and the **blue portion** of the bar represents the contribution of the the synthetic portion. Models tend to have higher performance on real images (larger green area) than on synthetic ones (smaller blue area). However, the difference in performance on the two subsets decreases as overall accuracy (bar height) decreases, with the DeepSeek-VL2 family, the PerceptionLM family, Llama 3.2 Instruct 11B, and Llava-OneVision 7B, having only a small difference between the two subsets.

4.4 Models trained on multi-image inputs show improved ability to reason across scenes

Finally, we explore which levers offer promise for advancing multimodal models' capacity to reason across scenes. We analyze performance based on whether models are explicitly trained on multi-image inputs, with CoT reasoning, and at large scale (many model parameters) in Figure 6. We find that CoT reasoning, which unlocks "thinking" tokens to parse scenes, has a mixed effect on reasoning across scenes, despite boosting single image perception across both model families we studied (78% for DeepSeek versus 70% for Qwen and 77% LlamaV-o1 versus Llama 3.2 Instruct 73%). This suggests standard reward based reasoning requires further research to enable reasoning across scenes. On the other hand, we see promise in models trained with multi-image inputs have 3× higher accuracy on `Common-O Bench` compared to those trained with single image training. We also, perhaps unsurprisingly found that larger models had stronger performance, which suggests that scaling model size may help boost accuracy.

(a) Reasoning Models (CoT) (b) Average Multi-image training (c) Average model scale

Figure 6: **Accuracy on our benchmark in different settings:** In (a), model families differ in whether their reasoning models (with CoT) perform better or works on `Common-O Bench`. In (b) and (c), we see improved reasoning for models that utilized multi-image training and were larger overall, suggesting using two approaches may enable better performance on `Common-O Bench`. Note: We average across several models when they have the same size or training-setup.

5 Discussion

Limitations. The real images in our benchmark were all taken by the authors, which understandably may reflect some bias in terms of locations, backgrounds, and objects used. The usage of synthetic image helps in some sense for more image diversity. Additionally, multiple choice setups are known to be somewhat brittle (Zheng et al., 2023; Long et al., 2024; Gupta et al., 2025)—simple changes to prompts and the order of choices can impact performance. An ideal setting would be open-ended generation, where models are able to use describe and reason about objects with their own labels. We also only include English text. Multilingual evaluation settings could be interesting future work, as models are increasingly trained in multiple languages.

Contributions With multimodal models saturating vision leaderboards focused on perception, we introduce `Common-O Bench`, a challenge for reasoning across scenes. We find while perceiving objects in a single image is easy, reasoning across the same scenes is challenging: the best performing model reaches just 35% accuracy on `Common-O Bench`– and no model is above 1% on our challenging subset `Common-O Complex`. We discover models are prone to hallucination when similar objects are present suggesting models may still be relying on object co-occurrence seen during training. To advance the essential skill of reasoning across scenes, new training paradigms that explicitly incorporate multi-image inputs with forms of reasoning beyond existing reward feedback are called for to overcome the challenge of hallucinations when reasoning across scenes.

Acknowledgements We thank Olga Golovneva, Kamalika Chaudhuri and Christoph Feichtenhofer for their thoughtful feedback on our paper and for suggesting exciting experiments based on our findings.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022.

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. UniBench: Visual Reasoning Requires Rethinking Vision-Language Beyond Scaling, August 2024. URL `http://arxiv.org/abs/2408.04810`. arXiv:2408.04810 [cs].

Alan Baddeley. The magical number seven: Still magic after all these years? In *Psychological Review*, volume 101, pp. 353—-356, 1994. doi: https://doi.org/10.1037/0033-295X.101.2.353.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.

Nelson Cowan, Candice Morey, and Zhijian Chen. The legend of the magical number seven. *Tall tales about the brain: Things we think we know about the mind, but ain't so*, pp. 45–59, 2007.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, Sean Christofferson, James Fort, Xiaqing Pan, Mingfei Yan, Jiajun Wu, Carl Yuheng Ren, and Richard Newcombe. Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SUGARCREPE++ Dataset: Vision-Language Model Sensitivity to Semantic and Lexical Alterations, June 2024. URL `http://arxiv.org/abs/2406.11171`. arXiv:2406.11171 [cs].

EpicGames. URL `https://www.unrealengine.com`. Unreal Engine is a copyright of Epic Games, Inc. and its affiliates (collectively, "Epic"). Any use of images, datasets, or other content made available by Epic, including without limitation through the Unreal Engine Marketplace or the Epic Games Launcher, in connection with your use of the PUG environments and datasets we've outlined in this paper and released publicly in connection hereto (the "PUG environments and datasets") or otherwise, is subject to the Epic Content License Agreement available at `https://www.unrealengine.com/en-US/eula/content` or other agreement between you and Epic.

Steven M. Frankland and Joshua D. Greene. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, 112(37): 11732–11737, 2015. doi: 10.1073/pnas.1421236112. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1421236112`.

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. *European Conference on Computer Vision*, 2024. doi: 10.48550/arXiv.2404.12390.

Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An Image is Worth More Than a Thousand Words: Towards Disentanglement in the Wild, October 2021. URL `http://arxiv.org/abs/2106.15610`. arXiv:2106.15610 [cs].

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, November 2022. URL `http://arxiv.org/abs/1811.12231`. arXiv:1811.12231 [cs, q-bio, stat].

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52733.2024.01363.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. Changing answer order can decrease MMLU accuracy. In *Workshop on Datasets and Evaluators of AI Safety*, 2025. URL `https://openreview.net/forum?id=MISIKTzC22`.

Arshia Hemmat, Adam Davies, Tom A Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in Plain Sight: Evaluating Abstract Shape Recognition in Vision-Language Models.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization, July 2021. URL `http://arxiv.org/abs/2006.16241`. arXiv:2006.16241 [cs, stat].

Hexiang Hu, Ishan Misra, and Laurens Van Der Maaten. Evaluating text-to-image matching using binary image selection (bison). In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. *arXiv preprint arXiv: 2502.11492*, 2025.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *arXiv:1612.06890 [cs]*, December 2016. URL `http://arxiv.org/abs/1612.06890`. arXiv: 1612.06890.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Koen Kraaijveld, Yifan Jiang, Kaixin Ma, and Filip Ilievski. COLUMBUS: Evaluating COgnitive Lateral Understanding through Multiple-choice reBUSes, December 2024. URL `http://arxiv.org/abs/2409.04053`. arXiv:2409.04053 [cs].

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3426–3440, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024a.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. VHELM: A Holistic Evaluation of Vision Language Models, October 2024b. URL `http://arxiv.org/abs/2410.07112`. arXiv:2410.07112.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning, 2023a. URL `https://arxiv.org/abs/2306.05425`.

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*, 2024a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024b.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial VQA: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2042–2051, 2021.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. *Conference on Empirical Methods in Natural Language Processing*, 2023b. doi: 10.48550/arXiv.2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, Zürich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Wei Lin, Muhammad Jehanzeb Mirza, Sivan Doveh, Rogerio Feris, Raja Giryes, Sepp Hochreiter, and Leonid Karlinsky. Comparison visual instruction tuning, 2024. URL https://arxiv.org/abs/2406.09240.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? *European Conference on Computer Vision*, 2023. doi: 10.48550/arXiv.2307.06281.

Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F Chen, and Min-Yen Kan. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. *arXiv preprint arXiv:2408.08656*, 2024.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv: 2408.02718*, 2024.

Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, March 1956. URL http://www.musanim.com/miller1956/.

Pranshu Pandya, Vatsal Gupta, Agney S. Talwarr, Tushar Kataria, Dan Roth, and Vivek Gupta. NTSEBENCH: Cognitive Reasoning Benchmark for Vision Language Models, April 2025. URL http://arxiv.org/abs/2407.10380. arXiv:2407.10380 [cs].

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision Language Models are blind. In Minsu Cho, Ivan Laptev, Du Tran, Angela Yao, and Hongbin Zha (eds.), *Computer Vision – ACCV 2024*, volume 15476, pp. 293–309. Springer Nature Singapore, Singapore, 2025. ISBN 978-981-9609-16-1 978-981-9609-17-8. doi: 10.1007/978-981-96-0917-8_17. URL https://link.springer.com/10.1007/978-981-96-0917-8_17. Series Title: Lecture Notes in Computer Science.

Megan Richards, Polina Kirichenko, Diane Bouchacourt, and Mark Ibrahim. Does Progress On Object Recognition Benchmarks Improve Real-World Generalization?, July 2023. URL http://arxiv.org/abs/2307.13136. arXiv:2307.13136 [cs].

Candace Ross, Melissa Hall, Adriana Romero-Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=LFfktMPAci.

Shailaja Keyur Sampat, Mutsumi Nakamura, Shankar Kailas, Kartik Aggarwal, Mandy Zhou, Yezhou Yang, and Chitta Baral. VL-GLUE: A Suite of Fundamental yet Challenging Visuo-Linguistic Reasoning Tasks, October 2024. URL http://arxiv.org/abs/2410.13666. arXiv:2410.13666 [cs].

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223, 2017.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644/.

Yixuan Sun, Yiwen Huang, Haijing Guo, Yuzhou Zhao, Runmin Wu, Yizhou Yu, Weifeng Ge, and Wenqiang Zhang. Misc210k: A large-scale dataset for multi-instance semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. Findings of the shared task on troll meme classification in tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pp. 126–132, 2021.

Hajime Takechi and Hiroko Dodge. Scenery picture memory test: A new type of quick and effective screening test to detect early stage alzheimer's disease patients. *Geriatrics & gerontology international*, 10:183–90, 04 2010. doi: 10.1111/j.1447-0594.2009.00576.x.

Qwen Team. Qvq: To see the world with wisdom, December 2024. URL https://qwenlm.github.io/blog/qvq-72b-preview/.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality, April 2022. URL http://arxiv.org/abs/2204.03162. arXiv:2204.03162 [cs].

Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-Producing Failures of Multimodal Systems with Language Models, March 2024a. URL http://arxiv.org/abs/2306.12105. arXiv:2306.12105 [cs].

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, April 2024b. URL http://arxiv.org/abs/2401.06209. arXiv:2401.06209 [cs].

Tomer Ullman. The Illusion-Illusion: Vision Language Models See Illusions Where There are None, December 2024. URL http://arxiv.org/abs/2412.18613. arXiv:2412.18613 [q-bio].

Zhecan Wang, Junzhang Liu, Chia-Wei Tang, Hani Alomari, Anushka Sivakumar, Rui Sun, Wenhao Li, Md Atabuzzaman, Hammad Ayyubi, Haoxuan You, Alvi Ishmam, Kai-Wei Chang, Shih-Fu Chang, and Chris Thomas. JourneyBench: A Challenging One-Stop Vision-Language Understanding Benchmark of Generated Images, January 2025. URL http://arxiv.org/abs/2409.12953. arXiv:2409.12953 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

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 preprint arXiv:2407.13766*, 2024a.

Xindi Wu, Hee Seung Hwang, Polina Kirichenko, and Olga Russakovsky. Compact: Compositional atomic-to-complex visual capability tuning. *arXiv preprint arXiv: 2504.21850*, 2025.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

Antonia Wüst, Tim Tobiasch, Lukas Helff, Inga Ibs, Wolfgang Stammer, Devendra S. Dhami, Constantin A. Rothkopf, and Kristian Kersting. Bongard in Wonderland: Visual Puzzles that Still Make AI Go Mad?, February 2025. URL http://arxiv.org/abs/2410.19546. arXiv:2410.19546 [cs].

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv: 2308.02490*, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5014–5022, 2016.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv: 2408.13257*, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All of the claims in our paper are supported by experiments conducted on 13 state-of-the-art vision language models. We run multiple analyses to validate our dataset and ground our contribution in prior work

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a limitations section where we discuss the only including English captions, the multiple-choice question setup. We also flag the potential for image bias due to having 4 photographers.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not introduce or use any theoreoms, formulas or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the model families and sizes, the hyperparameters we used and the inference libraries by model (*e.g.* vLLM, HuggingFace). All of the data is included with a hyperlink, to enable others to trace its provenance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide all of the data (images and CSVs with specific model input) used for our experiments in the recommended Croissant format.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We are an evaluation-only benchmark. We provide the hyperparameters we used for evaluation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We run statistical tests such as Pearson's correlation to validate trends we observe.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the compute resources we used, including specifying types of GPU and number of nodes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read through and followed the Code of Ethics for every portion of our research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our dataset is an evaluation-only benchmark and does not include any newly trained models. We do discuss that we were intentional in the data we chose to include (*e.g.* no images of people).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our contribution is an evaluation-only benchmark. We do not include any images of people. All images were manually gathered (either by the research team taking them, or by synthetically generating them using a game engine).

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The data introduced by our benchmark was introduced by the authors, all of whom provided consent for their images to be used for model evaluations and are credited for their contributions in the paper. The data license is included in our data repository, which is linked in the main paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This is an evaluation benchmark, where we provide all images, examples as CSVs, and a description of the dataset structure.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not perform any crowd sourcing or use any human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We did not use any study participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We do not leverage LLMs for any portion of our dataset creation or research design.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Model	Accuracy	Bootstrap Mean	Standard Error
Qwen 7B	65.16	65.21	0.02
Qwen 32B	70.34	70.35	0.02
Qwen 72B	69.18	69.17	0.02
Llava OneVision Chat 7B	78.17	69.73	0.02
Llava OneVision Chat 72B	77.7	77.17	0.02
DeepSeek VL Small	78.3	78.3	0.02
DeepSeek VL	69.8	69.8	0.02
LlamaV-o1 11B	77.28	77.27	0.02
Llama 3.2 Instruct 11B	73.42	73.42	0.02
Llama 4 Instruct Scout	80.31	80.33	0.02
Perception LM 3B	79.55	79.56	0.02
Perception LM 8B	84.59	84.60	0.02
GPT-4o	77.28	77.29	0.02

Table 3: Single image accuracy with standard error using bootstrapping with 1000 iterations.

A Image Taking Guidelines

We used the following procedure to guide our creation of images. First, each image taker selected a set of up to 7 objects and identified a background (e.g. a blanket, counter, or on the floor). Second, they take images iteratively, starting by placing a single object on the background and subsequently adding others (N=1 to N=7). Images were framed with the objects in the center or slightly off center (e.g. in Figure 2b, the plants in the third set of images from the left has leaves outside of the top part of the frame), with the goal that the majority if not the entirety of the object be contained within the frame. Across scenes, objects are often viewed from different viewpoints (e.g. top-down, versus side-view). Objects also may be partially occluded by other objects in the scene (e.g. in the bottom left image in Figure 2b the eye-mask is slightly occluded by the pink ball), but occlusions should be minimal with the restriction that all objects be easily human recognizable. For each scene (set of objects against a background), the image-taker would also take images from multiple visual orientations freely (with no restriction on the angle between the camera and the objects, so as to better capture real world diversity). Third, the image-taker would repeat against a new background, and add the objects to the scene in a different order and at a different orientation. Throughout this process, image-takers refrained from including any sensitive objects which may have privacy or IP concerns (e.g. humans, animals, brands, logos etc.) in images. Images were taken using smart phone cameras (Google Pixel, iPhone 15 Pro), as smart phones are one of the predominant modes of image creation currently.

B Additional Analysis

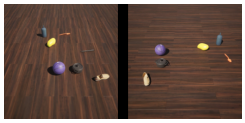
Accuracy for Single Image Perception Versus Multi-Image Reasoning Standard error We show in Table 3 the single image performance with standard error. We report the same performance for the multi-image reasoning task in Table 4. To compute the standard error we run bootstrapping with 1000 iterations on both the single image (baseline) and multi-image settings. Overall, we find a very small standard error.

Prompt Variants and Temperature We report two additional prompt reformulations (along with temperature ablations) for a total of 3 prompts on Qwen 7B. For a given temperature, we find the overall performance differs by 1.5-2.6% across prompts suggesting our claims are robust to prompt reformulations. We provide a full table of these results in Appendix B.

Role of Object Similarity In Table 6, we show the correlation between accuracy and the average similarity of objects in the scene. We observe a statistically significant negative correlation suggesting as models are more likely to make mistakes when objects are similar.

Additional model examples and mistakes In Figure 7, we show additional randomly sampled examples from `Common-O Bench`. In Appendix B, we show randomly selected mistakes in `Common-O Bench` across all models. The examples show the high degree to which models hallucination objects that are not in the ground truth.

(a) *answer:* glass cup, twine



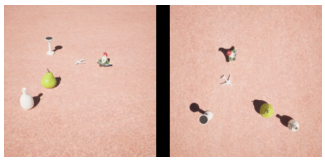
(b) *answer:* basketball, spoon, fakefruit, vase, shoes, cast iron

(c) *answer:* pink pot, pail with handle

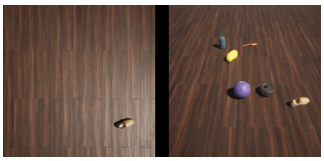
(d) *answer:* vase, shoes

(e) *answer:* No objects are in common

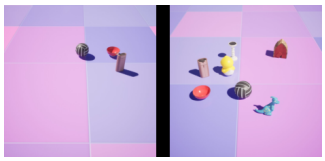
(f) *answer:* bottle opener



(g) *answer:* fakefruit, airplane, vase, orb, candle holder

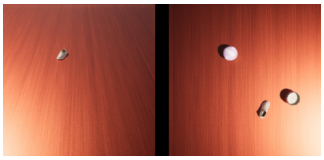


(h) *answer:* shoes



(i) *answer:* shaver, volleyball, bowl

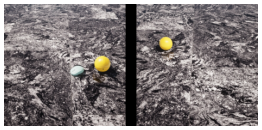
(j) *answer:* bottle opener



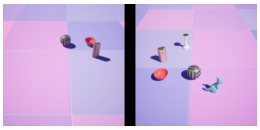
(k) *answer:* shoes

(l) *answer:* measuring tape

(m) *answer:* No objects are in common



(n) *answer:* volleyball



(o) *answer:* bowl, shaver, volleyball

(p) *answer:* kitchenware, basketball, football, cast iron

(q) *answer:* football, volleyball, keyboard, birdhouse, mouse

(r) *answer:* Stapler

Figure 7: Randomly sampled examples from `Common-O Bench`.

Model	Accuracy	Bootstrap Mean	Standard Error
Qwen 7B	13.26	13.27	0.01
Qwen 32B	24.16	24.14	0.01
Qwen 72B	18.53	18.55	0.01
Llava OneVision Chat 7B	2.61	2.61	0.005
Llava OneVision Chat 72B	14.84	14.85	0.01
DeepSeek VL Small	3.31	3.31	0.005
DeepSeek VL	3.06	3.07	0.005
LlamaV-o1 11B	8.9	8.9	0.009
Llama 3.2 Instruct 11B	5.07	5.08	0.006
Llama 4 Instruct Scout	35.12	35.14	0.01
Perception LM 3B	1.04	1.04	0.003
Perception LM 8B	1.86	1.86	0.004
GPT-4o	35.11	35.11	0.01

Table 4: Multi-image reasoning accuracy with standard error using bootstrapping with 1000 iterations.

Temperature	Prompt #1 Acc.	Prompt #2 Acc.	Prompt #3 Acc.
0.0	13.6	11.3	12.5
0.2	13.2	11.0	11.3
0.4	13.4	11.5	11.7
0.6	13.5	11.2	11.4
0.8	13.1	11.6	11.9
1.0	13.3	11.2	12.0

Table 5: We report accuracy across prompt reformulations across six temperatures for Qwen 7B.

C Synthetic data

The synthetic data was generated using Unreal Engine (EpicGames) and assets from Aria Digital Twins Catalog (Dong et al., 2025). We bought the following asset on fab to get the floor texture with a professional license: `https://www.fab.com/listings/66985cc5-13c2-45eb-9b5b-628ef4445a5c`. We randomly placed the assets into one of 16 different positions and apply some slight random rotation over the assets. To ensure that assets are not overlapping with each other, we constrained them to a given maximum size while keeping their aspect ratio. For each scene, we took images coming from 4 different camera positions.

D Statistical Analysis of Results

To get an approximation of the variance, we run bootstrapping with 1000 iterations on both the single image (baseline) and multi-image settings. The results are included below in Table . Overall, we find a very small standard error.

Model	Pearson Correlation
Qwen 7B	**-0.33***
Qwen 32B	**-0.38***
Qwen 72B	**-0.40***
Llava-OneVision Chat 7B	**-0.38***
Llava-OneVision Chat 72B	**-0.30***
DeepSeek-VL2 Small	-0.12*
DeepSeek-VL2	**-0.30***
LlamaV-o1 11B	-0.29*
LlamaV 3.2 11B	**-0.33***
Llama 4 Instruct Scout	**-0.41***
PerceptionLM 3B	-0.10
PerceptionLM 8B	**-0.35***

Table 6: Correlation between similarity among common objects and accuracy. The negative correlation shows that, the more similar the common objects are lead to lower accuracy. * indicates statistical significance with correlations of moderate strength or above in bold.

Model	*Choices (Enumerated by Letter to Model)*	*Ground Truth*	*Prediction*
GPT-4o	[silver grater, No objects are in common, dark chocolate bar wrapped in foil, silver straw, silver whisk, silver knife, tangerine, measuring cup]	No objects are in common	Measuring cup
Llava-OneVision	[No objects are in common, mallard (fake duck), vase, hammer, calculator, dish, basketball, fakefoodcan]	B, D, E	D, E, H
Qwen	[dumbbell, mouse, hammer, No objects are in common, football, birdhouse, keyboard, volleyball]	B, D, G	B
PerceptionLM	[spoon, No objects are in common, orange, glass, keys, lime, fork, popcorn kernel]	E	A, B, C
Qwen	[dino, candle holder, mallard (fake duck), bowl, volleyball, No objects are in common, shaver, birdhouse]	D, E, G	D
Llama 4 Instruct	[watermelon, plant, No objects are in common, coffee mug, earbuds, candle snuffer, pen, ball]	D, E, F, H	D
Llama 3.2 Instruct	[bottle opener, gold jigger, 2-prong serving fork, strainer, paring knife with wooden handle, No objects are in common, gold paring knife, silver jigger]	A, B, D, G, H	D, G
Llama 3.2 Instruct	[fakefruit, airplane, bowl, No objects are in common, spoon, football, keyboard, mouse]	C, F, G	D
Llama 3.2 Instruct	[fakefoodcan, vase, volleyball, spoon, kitchenware, No objects are in common, fakefruit, shoes]	A, B, E, G	A, B
Qwen	[remote, basketball, calculator, No objects are in common, mouse, vase, marker, volleyball]	C, E, H	C
Llama 3.2 Instruct	[fish bowl, white pill bottle, paint brush, candy cane, No objects are in common, orange pill bottle, lint roller, scissors]	B, F, H	B, D, F, G, H
PerceptionLM	[No objects are in common, candle, marker, fakefruit, keyboard, mallard (fake duck), bowl, remote]	B, C, E, G, H	A, B, C
GPT-4o	[cup, mallard (fake duck), vase, No objects are in common, football, candle, volleyball, shoes]	candle, shoes, vase, volleyball	shoes, volleyball
Llama 4 Instruct	[spoon, No objects are in common, fakefruit, cast iron, basketball, marker, vase, shoes]	C, D, G, H	C, D, H
Qwen	[spoon, cast iron, basketball, vase, fakefruit, No objects are in common, marker, shoes]	D, E, H	E, H
Qwen	[No objects are in common, fakefoodcan, fakefruit, shoes, spoon, vase, volleyball, kitchenware]	C, F, H	B, C, F
Llava-OneVision	[bowl, keyboard, No objects are in common, marker, remote, fakefruit, candle, mallard (fake duck)]	A, B, D, G	A, B, C
Qwen	[No objects are in common, pail with handle, burnt orange pot, leaf, black pot, easel, pink pot, watering can]	B, C, E, G	A
DeepSeekVL2	[No objects are in common, marker, basketball, calculator, vase, mouse, volleyball, remote]	B, D, E, F, G, H	A, B, C
Llava-OneVision Chat	[black pot, burnt orange pot, pink pot, pail with handle, No objects are in common, leaf, watering can, easel]	A, C	A, D, G

Table 7: Randomly sampled model mistakes in `Common-O Bench`.

Model	Multi Image			Single Image (Baseline)		
	Accuracy	*Bootstrap Mean*	*Std. Err*	*Accuracy*	*Bootstrap Mean*	*Std. Err*
Qwen 7B	13.26	13.27	0.01	65.16	65.21	0.02
Qwen 32B	24.16	24.14	0.01	70.34	70.35	0.02
Qwen 72B	18.53	18.55	0.01	69.18	69.17	0.02
Llava OneVision Chat 7B	2.61	2.61	0.005	78.17	69.73	0.02
Llava OneVision Chat 72B	14.84	14.85	0.01	77.7	77.17	0.02
DeepSeek VL Small	3.31	3.31	0.005	78.3	78.3	0.03
DeepSeek VL	3.06	3.07	0.005	69.8	69.8	0.02
LlamaV-o1 11B	8.9	8.9	0.009	77.28	77.27	0.02
Llama 3.2 Instruct 11B	5.07	5.08	0.006	73.42	73.42	0.02
Llama 4 Instruct Scout	35.12	35.14	0.01	80.31	80.33	0.02
PerceptionLM 3B	1.04	1.04	0.003	79.55	79.56	0.02
PerceptionLM 8B	1.86	1.986	0.004	84.59	84.60	0.02
GPT-4o	35.11	35.11	0.01	77.25	77.29	0.02

Table 8: Results of running bootstrapping with 1000 iterations. We show the average performance (`"Accuracy"`) versus the bootstrap mean and standard error on `Common-O Bench` and the single image baseline experiments.

E Dataset Card

We include a datasheet for `Common-O Bench` below, following the example from Gebru et al. (2021).

Motivation

For what purpose was the dataset created? The dataset was created the test the reasoning abilities of multimodal LLMs in multi-image, multi-object settings.

Who created the dataset? This is redacted during the review process to maintain anonymity and will be included in the camera-ready.

Who funded the dataset creation? This is redacted during the review process to maintain anonymity and will be included in the camera-ready.

Any other comments? None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. Each instance is a tuple of 2 images, a set of potential objects that are in both images and a set of the ground-truth, common objects between both images.

How many instances are there in total (of each type, if appropriate)? There are 10586 instances in `Common-O Bench` and 12600 instances in `Common-O Complex`.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? These were manually created instances, either via the authors taking the images or the authors using a game engine to synthetically create the images. We created a large set of synthetic images (≈400k). For `Common-O Bench` (N=3 to N=7 objects) and `Common-O Complex` (N=3 to N=7 objects), we randomly sampled images with the target number of objects.

Is there a label or target associated with each instance? The target associated with each instance is the set of objects in common between both images (*e.g.* apple, keys).

Is any information missing from individual instances? All of the information is included for every instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. Each image in a given contains a specific configuration of objects. This configuration is taken from multiple orientations. These orientations are labeled in the data files. Additionally, each image is contained with multiple instances. The instances in the data file are label with the image filenames so it's clear to see which instances have the same images.

Are there recommended data splits (e.g., training, development/validation, testing)? This is an evaluation-only benchmark; we do not provide any training or validation splits.

Are there any errors, sources of noise, or redundancies in the dataset? The instances were manually created. Potential sources of noise may come from ambiguitiy in idenitiying objects, which is captured by our human baseline.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is entirely self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? The dataset does not contain any confidential or private information.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships,

or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? The dataset does not contain any sensitive information.

Any other comments? None.

Collection Process

How was the data associated with each instance acquired? Every real photo was manually taken by one of the authors on this paper specifically for this dataset. Every synthetic photo was generated by the authors using a game engine. We manually wrote the set of objects found in each image.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software, programs, software APIs)? We used manual human curation for the real images and the Unreal engine for synthetic images. We validated the images by sampling a subset to hand-annotate.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

For the synthetic images, we manually downsampled via random sampling.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The authors performed all components of the data collection. We will include full details about the authors in the camera ready to preserve anonymity.

Over what timeframe was the data collected? The data was collected over about 3 months.

Were any ethical review processes conducted (e.g., by an institutional review board)? The data collection went through IRB. We did not include humans in the images.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The data was not collected from external individuals, third parties or web sources. We manually collected all data.

Were the individuals in question notified about the data collection? N/A; see previous question.

Did the individuals in question consent to the collection and use of their data? N/A; see previous question.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. N/A.

Any other comments? None.

Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section

We manually collected/generated all dataset instances and therefore did not perform any additional data processing beyond image resizing. All images in their original size were saved.

Uses

Has the dataset been used for any tasks already? The dataset has not been publicly released yet (outside of the private repository for paper review) and therefore has not been used for any additional tasks.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. The dataset is assessible through Kaggle at this link.

What (other) tasks could the dataset be used for? `Common-O Bench` has been tested for multiple-choice QA with multiple possible answers. The dataset could also be tested in open-ended question answering.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? There is very minimal risk for harm. We did not include any pictures of people, real or generated, and we also excluded any logos. Additionally, this dataset is only for evaluation and therefore will not be used in model training.

Are there tasks for which the dataset should not be used? The dataset is exclusively for evaluation and should not be used to train or finetune any models.

Any other comments? None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. Yes, the dataset will be publicly available on HuggingFace.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? We will host the dataset on HuggingFace. Because this paper is the introduction of the dataset, we will use the paper DOI.

When will the dataset be distributed? The dataset will be distributed upon acceptance of the paper in 2025.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The dataset is being distributed under the non-commercial CC BY-NC 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

Any other comments? None.

Maintenance

Who will be supporting/hosting/maintaining the dataset? `REDACTED AUTHORS` will be maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? `REDACTED AUTHORS` can be contacted through the email addresses provided in the camera ready.

Is there an erratum? If so, please provide a link or other access point. There is currently not an erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? We will update the dataset for any errors. We will likely communicate this via social media and perhaps a GitHub page.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers. N/A

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. We encourage anyone interested in potential augmentations and contributions to contact us using our email addresses, listed above.

Any other comments? None.