MEBench: A Novel Benchmark for Understanding Mutual Exclusivity Bias in Vision-Language Models

Anh Thai[1] Stefan Stojanov[1] Zixuan Huang[2] Bikram Boote[2] James M. Rehg[2]
[1]Georgia Institute of Technology, [2]University of Illinois, Urbana-Champaign

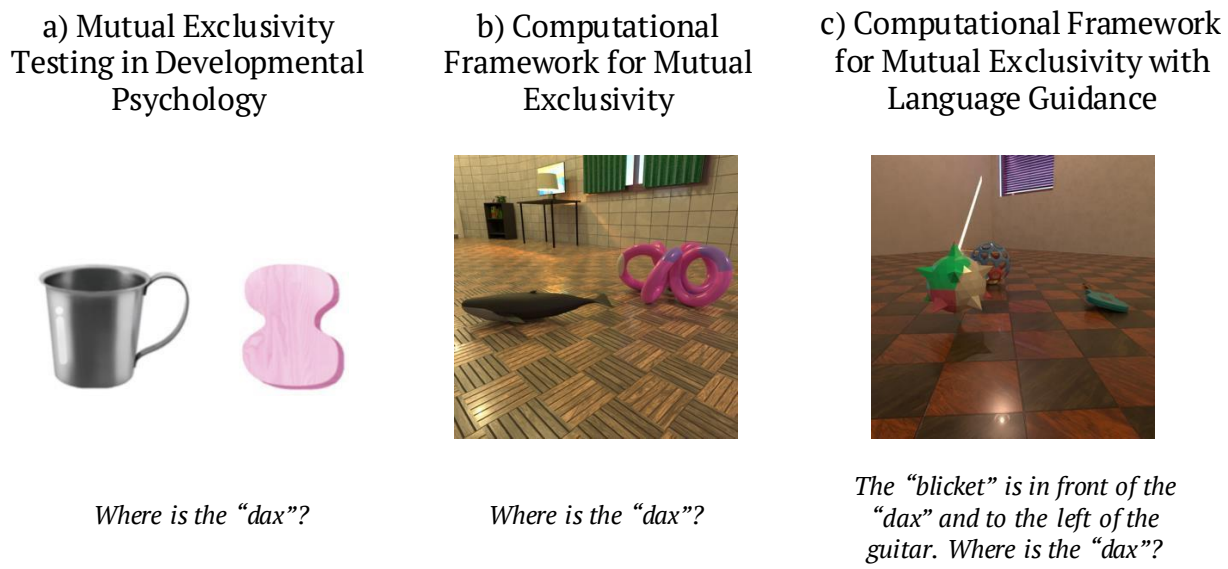


Figure 1. **Mutual Exclusivity Bias Evaluation Settings.** (a) Traditional ME bias evaluation in developmental psychology and early computational studies [6], (b) MEBench setup for classic ME bias testing, and (c) MEBench setup for evaluating ME bias in conjunction with spatial reasoning.

Abstract

This paper introduces MEBench, a novel benchmark for evaluating mutual exclusivity (ME) bias, a cognitive phenomenon observed in children during word learning. Unlike traditional ME tasks, MEBench further incorporates spatial reasoning to create more challenging and realistic evaluation settings. To facilitate controlled experimentation, we also present a flexible and scalable data generation pipeline that supports the construction of diverse annotated scenes. We assess the performance of various vision-language models (VLMs) on this benchmark using novel evaluation metrics that capture key aspects of ME-based reasoning. We find that these VLMs exhibit weak ME bias, while showing some ability to leverage extra spatial context to resolve ambiguity in multiple novel object settings. Project page: http://mebench.github.io/.

1. Introduction

During a developmental phase known as the vocabulary spurt [18], toddlers remarkably acquire 10–20 new words per week. This rapid word learning is supported by powerful inductive biases that help children resolve the ambiguity of mapping spoken words to real-world referents, referred to as object-label mapping. One such bias, mutual exclusivity, is the assumption that each object has a unique label. When presented with a mix of familiar and unfamiliar objects, children tend to associate novel words with the unknown items. For instance, if a caregiver says, "Look at the blicket," while the child sees her favorite Ducky, a Mickey Mouse figure, and a novel toy, mutual exclusivity guides the child to infer that "blicket" refers to the toy whose name is unknown.

Inspired by such cognitive mechanisms, researchers in computer vision have begun exploring how inductive biases can be incorporated into learning models to improve generalization and interpretability. Among these, shape bias,

the tendency to associate category membership with object shape, has received significant attention [20, 22, 30], particularly due to its relevance in 3D object categorization. In contrast, mutual exclusivity bias, despite being a foundational aspect of human language learning [6, 16, 17], remains relatively underexplored in computational settings.

With the rise of Vision-Language Models (VLMs) that aim to mirror human-like reasoning by grounding language in visual inputs, there is now an opportunity to study mutual exclusivity within a computational framework. These models provide a natural platform for investigating whether and how such biases emerge, especially in complex visual scenes that require reasoning.

In this work, we introduce MEBench, a new benchmark for evaluating mutual exclusivity bias in VLMs. Inspired by classic studies in developmental psychology, MEBench formulates the problem as an inference task: given an image containing both known and novel objects, along with a novel label, the model must (1) identify and localize the known categories and (2) associate the novel label with the correct unknown instance by leveraging mutual exclusivity. MEBench goes beyond prior work that focuses solely on object-label association [33] by introducing realistic, cluttered scenes and additional reasoning requirements. Specifically, models must not only detect and differentiate between known and novel objects but also apply spatial reasoning and contextual cues to resolve ambiguity when multiple unfamiliar items are present.

Since no real-world dataset exists for this task, we develop a scalable synthetic data generation pipeline. This pipeline transforms any 3D dataset of object categories into realistic scenes and includes a curated set of procedurally generated novel objects, using Blender's geometry nodes [1]. To prevent lexical leakage, novel object labels are sampled from synthetic, non-English-like words (e.g., "dax," "toma") commonly used in human studies of mutual exclusivity [29].

We systematically evaluate state-of-the-art VLMs on MEBench and analyze their performance across multiple subtasks, revealing both capabilities and limitations in their reasoning behavior. Mastery of MEBench would mark a meaningful step toward equipping AI systems with robust zero-shot generalization capabilities, essential for adaptive applications in real-world domains such as home robotics and human-AI interaction.

In summary, our contributions are:

1. We introduce MEBench, a computational framework for evaluating mutual exclusivity (ME) bias, and extend its evaluation beyond traditional settings by incorporating spatial reasoning with contextual scene descriptions.
2. We develop a flexible synthetic data generation pipeline that enables controlled experimentation and systematic analysis of factors affecting ME-based learning.
3. We benchmark a diverse set of vision-language models (VLMs) on MEBench, and find that modest increases in scene complexity cause a cliff in ME performance for all models, and leveraging spatial context improves performance in ambiguous multi-novel settings.

2. Related Work

Computational Mutual Exclusivity Modeling. The work most closely related to ours is [33], which introduces a generalized low-shot object learning framework that requires applying the mutual exclusivity assumption to associate a novel label with the correct novel object. While our approach also relies on mutual exclusivity, it differs in two key aspects. First, we incorporate a curated set of procedurally generated novel objects paired with pseudo words as labels, minimizing lexical leakage during evaluation. Second, we extend the task beyond traditional mutual exclusivity by introducing a setting with multiple novel objects, where additional reasoning, such as interpreting spatial cues, is required to resolve ambiguity. This level of disambiguation is not addressed in [33]. Another prior work, [6] demonstrated that existing feed-forward deep models fail at ME, but did not provide a comprehensive framing or benchmarking SOTA methods.

Other recent works have investigated various learning scenarios to explore the strategies children employ when acquiring new concepts [3, 8, 10]. We complement these efforts by introducing a comprehensive benchmark that studies the mutual exclusivity bias commonly observed in infants during the initial stages of word learning.

Language-guided Object Detection/Segmentation Tasks. Open-category object detection and segmentation [12, 38] settings require learning models to detect and segment both known and novel objects—novel to the localization models, though not necessarily to the LLMs—by prompting the LLMs. In contrast, our task focuses on learning object-label mappings, where labels may fall outside typical pretraining vocabularies, such as pseudo-names like "dax" commonly used in psychology studies [29].

Referring and reasoning segmentation [13, 24, 39] tasks require models to segment objects in an image based either on explicit natural language attribute-based descriptions or functional reasoning. Our setting extends beyond these tasks by requiring models to leverage ME bias to associate a novel label (a pseudo-name unfamiliar to the LLMs) with the novel object. When multiple novel objects are present, the model must further reason about spatial relationships to resolve ambiguity. Our approach emphasizes inter-object relationships, requiring the model to develop a holistic scene understanding alongside object-centric reasoning.

Synthetic 3D Datasets and Generators. Advances in realistic rendering engines [5, 7, 28] and image generation models [26, 27] have helped narrow the sim-to-real gap, improving generalization to real-world scenarios, and have been widely employed to study various tasks [7, 28, 30,

31, 36] where large-scale real-world data is unavailable. In our work, we use synthetic data to address a novel problem where real-world datasets do not yet exist. This approach enables the creation of diverse, realistic, and controllable environments for scalable experimentation.

Vision-Language Detection/Segmentation Models. With advancements in large vision-language models (VLMs), the ability of AI systems to comprehend images and respond to language prompts related to visual content has improved significantly [4, 14, 15, 19, 32] Beyond general image understanding, other works [24, 35, 37, 39, 40] focus on object grounding, which involves localizing and reasoning about objects within an image and outputting their locations, rather than solely providing text-based responses. These models require a precise scene representation to accurately ground objects based on language prompts. In this work, we use these models as our baselines for our MEBench due to their strong object detection capabilities and their ability to reason effectively with language inputs.

3. MEBench

In this section, we first formulate the task in Section 3.1, followed by a detailed description of the data generation pipeline in Section 3.2. Finally, we introduce the various data variants used in our study in Section 3.3.

3.1. Task Formalization

Given an RGB image containing multiple objects—some from known everyday categories and one or more novel objects—the model is asked to localize the referent of a novel label (e.g., "Where is the blicket?"). The task involves three subtasks: (1) **Object Detection**, further divided into: *Object Localization*—Identifying and localizing all objects present in the scene; and *Open-world Recognition*— Differentiating between known and novel instances using visual information; (2) **Novel Label Assignment**—Applying the mutual exclusivity assumption to correctly associate the novel label with the novel object.

When multiple novel objects are present, mutual exclusivity alone is insufficient to resolve ambiguity. To address this, we introduce spatial descriptions (e.g., "The dog is to the right of the blicket") that encode inter-object relations. The model must then perform: (3) **Spatial Reasoning**—the ability to interpret spatial descriptions in text to facilitate accurate novel label assignment. Note that this additional input can also be incorporated into the single unknown object scenario.

3.2. Data Generation Pipeline

Our system is designed to be compatible with any dataset of 3D object categories. For our benchmark, we use the Toys4K [30] dataset, chosen for its large number of categories (105) and diverse toy-like object appearances, which closely resemble the variability encountered in real-world child learning scenarios. This dataset comprises common, everyday object categories (e.g., "car", "dog"), which serve as the known categories in our benchmark.

Novel Objects. We introduce a curated set of 64 novel objects (Figure 3), manually designed and procedurally generated using geometric nodes in Blender [1], ensuring unique and diverse structures. The scale of our novel object set is comparable to prior benchmarks that evaluate generalization to unseen or uncommon categories (e.g., Bogard-HOI [11] considers 16–64 unseen classes, and other work [21] benchmarks VLMs on 50 uncommon objects), while the developmental psychology NOUN database [9] contains 45–64 objects with unusual names. This set is specifically crafted to minimize lexical leakage when paired with pseudo-word labels: foundation models trained on web-scale data have likely learned associations between common objects and their names, but are unlikely to have learned stable name–visual mappings for these synthetic, abstract, non-semantic shapes. To further ensure fairness, we randomize the pairing between novel objects and pseudo-word labels at inference time, preventing any fixed object–word assignment from being exploited.

Background. To create diverse and realistic backgrounds, we generate room environments as the backdrop for our scenes. These rooms are primarily living rooms and bedrooms, as they represent the most natural settings for child play. Our room generation is based on Infinigen [23], with each room including varied background object configurations. Lighting conditions are naturally derived from indoor sources or outdoor light filtering through open doors and windows (see Figure 2).

Data Rendering. During each scene rendering, we first randomly select a subset of objects from the known categories and a set of novel objects from our curated collection. The total number of objects and the number of novel instances are determined based on user-defined input parameters. To achieve realistic object placement, we use rigid body simulation to generate natural rotational poses. Objects are then scaled and positioned at random locations within the scene while ensuring that no two objects collide.

Scene Description Generation. We generate scene descriptions as additional contextual inputs for each view of the scene, rather than for the entire 3D scene. This is because some spatial relationships in 3D are inherently viewpoint-dependent terms like "left," "right," "in front of," and "behind" can vary depending on the observer's perspective. For each view, we first construct a scene graph using object bounding boxes and the corresponding depth map. In this graph, objects are represented as nodes, while pairwise spatial relationships form the edges. We then translate the structured scene graph representation into plain English descriptions. For instance, given the scene graph expression:

Figure 2. **Example of Rendered Data for the MEBench Benchmark.** We systematically generate diverse object configurations within varied room backgrounds, ensuring photorealistic renderings that capture realistic spatial arrangements and lighting conditions.

Figure 3. **Novel Objects in MEBench.** To minimize lexical leakage during evaluation, we constructed a database of novel objects using procedural generation in Blender [1] with geometry nodes from GeoShapeV2 [2] and Thingi10K [41], paired with pseudo-words as labels.

"Dog":{"to left of":["dax", "pig"]} we generate the natural language description: "The dog is to the left of the dax and the pig." Notably, this scene description generation process is fully deterministic, ensuring that all pairwise object relationships are consistently included in the description. This eliminates potential ambiguities and provides a structured yet flexible input format for downstream reasoning tasks.

View-point Selection for Inference. During inference, we select rendered viewpoints where all objects in the scene are visible. An object is considered visible in a given view if its segmentation mask occupies at least 200 pixels in a 224×224 resolution image. To further minimize variability in object visibility, we run each model three times, each with a different set of selected viewpoints, and a different novel object–pseudo-word assignment.

3.3. Data Variants

To enable a comprehensive analysis and gain deeper insights into the limitations of different methods, we generate datasets with progressively increasing levels of difficulty. Each variant is specifically designed to evaluate three key model capabilities: (1) Object Localization, (2) Mutual Exclusivity (ME) Bias, and (3) Spatial Reasoning. These capabilities are assessed along two primary axes: Number of known objects and Number of novel objects. To maintain clarity, we define data variant acronyms based on these axes. For example, 1K-0U represents a scenario with 1 known object and 0 novel objects. These data variants provide

Table 1. **Overview of Data Variants for Evaluating VLM Baselines.** The dataset variants are categorized into three groups, each designed to assess a specific subtask: object localization, mutual exclusivity (ME) bias, and spatial reasoning ability.

Setting	Data Axis		Model Assessment		
	#Known Obj.	#Novel Obj.	Obj. Localization	ME Bias	Spatial Reasoning
1K-0U	1	0	✓	✗	✗
1K-1U	1	1	✓	✓	✗
2K-1U	2	1	✓	✓	✗
1K-2U	1	2	✓	✓	✓

a framework for systematically testing model capabilities, distinguishing basic object detection from higher-level reasoning tasks such as ME bias and spatial inference (Table 1).

Object Localization: Evaluating Detection Ability. The first group consists of only one known object, without any novel objects: 1K-0U. This setting assesses the model's ability to localize known objects. A model that performs well in this group demonstrates object detection capabilities, independent of novel object inference.

Mutual Exclusivity Bias: Assigning Novel Labels Correctly. The second group introduces one novel object alongside known objects (1K-1U and 2K-1U). This group evaluates the model's ability to apply the ME assumption, which assigns the novel label to the novel object.

Spatial Reasoning: Disambiguating Multiple Novel Objects. The third group contains two novel objects: 1K-2U, introducing inherent ambiguity that cannot be resolved solely through ME bias. This setting evaluates the model's ability to leverage spatial descriptions to disambiguate novel objects based on their relationships within the scene.

Due to the procedural nature of our data generation pipeline, the benchmark can be readily scaled to more challenging settings, such as scenes with additional known and novel objects or more cluttered backgrounds. In this work, however, we focus on systematically analyzing model behavior across the core evaluation axes above, rather than substantially increasing scene complexity.

4. Experiments

4.1. Baselines

We evaluate our MEBench benchmark on SOTA VLM baselines, consisting of both closed-source and open-source models: CogVLM [35], Gemini 2.0 [32], Sa2VA [39], OMG-LLaVA [40], LISA [13], LLaVA-OV [14], and F-LMM [37]. Most of these VLMs are trained primarily for text-guided object grounding, such as detecting or segmenting objects based on language prompts, which makes them particularly relevant baselines for our task given its close connection to object detection. In contrast, LLaVA-OV [14] is designed as a more general VQA and reasoning model. The diversity in model architectures and training objectives allows us to conduct a comprehensive analysis, gaining deeper insights into how different VLMs perform on our benchmark across various reasoning and perception challenges. For all experiments, we evaluate each baseline three times, using a different subset of scene viewpoints and a different novel object–pseudo-word assignment in each run. We observe that performance is highly consistent across runs, with negligible standard deviation for all models (Figure 4).

4.2. Data

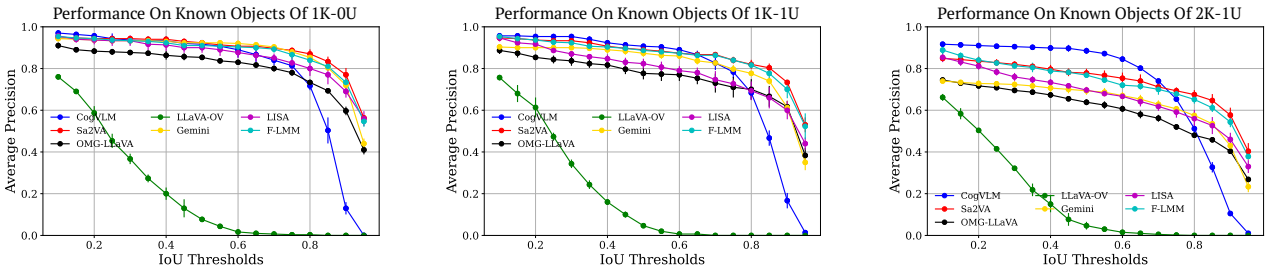
Using the data generation pipeline from Section 3.2, we generate 100 scenes per setting. This number provides substantial variation in object combinations, configurations, backgrounds, and lighting conditions, while keeping the dataset manageable for evaluation. Additionally, each scene is rendered from 25 randomly sampled viewpoints, ensuring diverse viewing perspectives. In total we achieve 2,500 images per setting, which provides robust coverage of object appearances, spatial arrangements and viewpoints. During inference, we ensure that all objects are visible in the selected views to maintain consistency and fairness in evaluation (see Section 3.2).

4.3. Evaluation Protocol & Metrics

In this subsection, we outline the evaluation criteria for each subtask and the metrics used to assess model performance.
Object Localization. We assess the models' ability to detect objects using the standard Average Precision at IoU threshold ($AP@t$) metric, commonly used in object detection tasks [25, 34].
Mutual Exclusivity Bias. Mutual exclusivity (ME) bias refers to the tendency to assign a novel label to a novel object, rather than to an already familiar one. To assess whether a model exhibits ME bias, we evaluate its ability to assign the pseudo-word label to the correct novel object, conditioned on successfully recognizing all familiar objects present in the scene. This conditioning is important because it excludes cases where the model fails to recognize a known object and therefore cannot meaningfully apply mutual exclusivity. We first define $p(x_{n\to n})$ as the probability of correctly assigning the novel label to a novel object, equivalent to the ME score from [6], and conceptually similar to $AP@0.5$ metric for novel objects. Similarly, $p(x_{k\to k})$ denotes correct assignment of known labels, and $p(x_{n\to k})$ captures incorrect assignment of the novel label to a known object. We focus on conditional metrics given correct known-object classification: $p(x_{n\to n}|x_{k\to k})$: correct novel label assignment; and $p(x_{n\to k}|x_{k\to k})$: incorrect novel label assignment to a known object, introducing a normalized ME score:

$$-1 \leq \text{ME} = \frac{p(x_{n\to n}|x_{k\to k}) - p(x_{n\to k}|x_{k\to k})}{p(x_{n\to n}|x_{k\to k}) + p(x_{n\to k}|x_{k\to k})} \leq 1$$

The ME score is defined as: (1) $-1 \leq ME \leq 0$: Tendency to misassign the novel label to a known object, which indicates weak ME bias; (2) $0 < ME \leq 1$: Tendency to assign the novel label correctly. Higher ME score indicates stronger ME bias. In addition, the model may: (1) Refuse to predict on novel labels: $p(x_{n\to\varnothing} \mid x_{k\to k})$, indicating conservative behavior; (2) Misclassify background objects as novel or fail to accurately segment the objects, leading to low IoU: $p(x_{n\to bg} \mid x_{k\to k})$, often due to poor object detection.



(a) (b) (c)

Figure 4. **Object Detection Performance of VLMs on Known Objects** in the (a) 1K-0U (1 known and 0 unknown object), (b) 1K-1U (1 known and 1 unknown objects), and (c) 2K-1U (2 known and 1 unknown objects. Each baseline is run three times and performance's standard deviation is shown as vertical bar at each data point.

Spatial Reasoning. To evaluate the model's spatial reasoning ability, we analyze the improvement in the probability of correctly assigning the novel label to the intended novel object, alongside the reduction in misclassifications among novel objects. We quantify the impact of scene spatial descriptions on model performance by measuring the proportional increase in correctly assigning novel labels to novel objects when spatial context is provided compared to when it is absent. Specifically,

$$\text{Spatial Reasoning} = \frac{p_w(x_{n\to n}|x_{k\to k}) - p_{w/o}(x_{n\to n}|x_{k\to k})}{p_{w/o}(x_{n\to n}|x_{k\to k})}$$

where $p_w(.)$ and $p_{w/o}(.)$ represent the performance with and without spatial input, respectively. A positive Spatial Reasoning score indicates that the model effectively uses spatial context to enhance its performance, demonstrating a strong spatial reasoning capability.

When the scene contains multiple unknown objects, we additionally evaluate the ambiguity score, which quantifies the error rate of incorrectly assigning the novel label to the wrong novel object. This metric provides insight into the model's ability to disambiguate between multiple novel objects using available contextual cues.

$$\text{Ambiguity} = \frac{p(x_{n\to no}|x_{k\to k})}{p(x_{n\to n}|x_{k\to k}) + p(x_{n\to no}|x_{k\to k})}$$

where $p(x_{n\to no}|x_{k\to k})$ represents the probability of misassigning the novel label to the incorrect novel object, given that the known objects have been correctly identified.

These metrics generalize to scenes with more objects. In multi-novel-object settings, $p(x_{n\to n} \mid x_{k\to k})$ captures both correct and incorrect assignments among novel objects, enabling broader evaluation of a model's ability to manage complexity and ambiguity.

5. Results

5.1. Object Detection

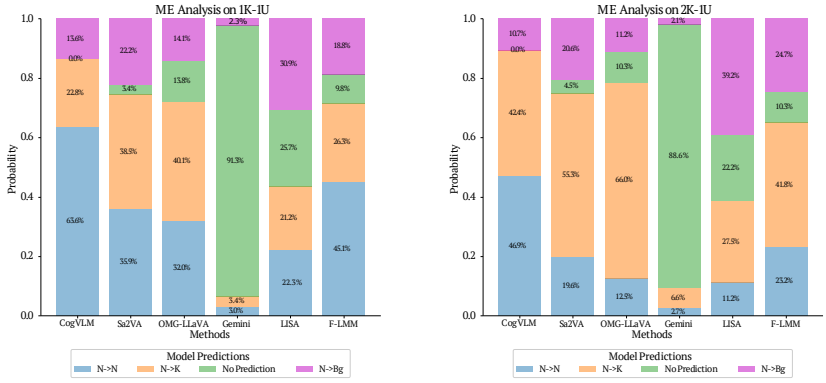
High Performance on Known Objects. In Figure 4, we show the performance of various baselines in the 1K-0U setting. Except for LLaVA-OV—primarily designed for general VQA and reasoning rather than object detection or grounding—most methods achieve consistently high performance. When an unknown object is introduced (Figure 4b), performance shows only a slight decline across all methods, indicating strong robustness in detecting common, known objects.

Extra Known Object Decreases Performance. In Figure 4c, we observe a significant performance drop across most models when an additional known object is introduced into the scene. This decline can be attributed to increased clutter, occlusion, and potential confusion between similar known objects. However, CogVLM [35] demonstrates greater robustness in object detection across both settings, effectively distinguishing objects despite the added complexity.

5.2. VLMs Show Weak ME Bias

In this subsection, we examine the ME bias of baseline models by analyzing the distribution of their responses to novel labels, specifically in the case where they have correctly identified the known object(s). This analysis provides insights into how effectively each model applies the ME assumption when associating novel labels with novel objects. In these settings we set the threshold t in the $AP@t$ metric to 0.5 for evaluation.

In Figures 5 (left and middle), we present the response distributions of six baseline models, along with their ME scores (Figure 5, right), computed using the formula introduced in Section 4.3. Among all baselines, CogVLM [35] achieves the highest ME score, indicating a strong mutual exclusivity bias. F-LMM [37] and LISA [13] follow, though with a significant performance gap. The remaining methods exhibit negative ME scores, suggesting a weaker ME bias,



Method	1K-1U ↑	2K-1U ↑
CogVLM	0.472	0.050
F-LMM	0.263	-0.286
LISA	0.025	-0.421
Sa2VA	-0.035	-0.477
Gemini	-0.063	-0.419
OMG-LLaVA	-0.112	-0.682

Figure 5. **Mutual Exclusivity (ME) Analysis in the (Left) 1K-1U and (Middle) 2K-1U settings**. These settings contain one novel object in the scene. The response types are categorized as follows: $N \rightarrow N$ denotes correctly assigning the novel label to the novel object, $N \rightarrow K$ represents misassigning the novel label to a known object, and $N \rightarrow Bg$ indicates misassigning the novel label to a background distractor or failing to detect high-quality object bounding boxes. Additionally, No Prediction indicates cases where the model fails to produce a bounding box for the referred object. **(Right) ME Scores** of 1K-1U and 2K-1U settings. Higher scores indicate stronger ME bias.

as they frequently misassign novel labels to known objects, failing to leverage ME. Introducing an additional known object leads to a notable drop in the ME scores. Specifically, we observe a significant decrease in correctly assigning the novel label to the novel object, and a corresponding increase in misassigning the novel label to known objects, as reflected in the blue and orange segments of the bar charts. This suggests that introducing an additional known object into the scene substantially increases task difficulty, leading to a significant decline in ME scores ($\Delta_{avg} = 0.464$). Additionally, the scene becomes more cluttered and occluded, further challenging the model's ability to accurately distinguish between known and novel objects.

Notably, Gemini [32] has the highest rate of missing predictions for novel objects, failing to produce a bounding box nearly 90% of the time. This behavior can be attributed to its training paradigm, which prioritizes reducing hallucinations in general conversations. In contrast, CogVLM [35] consistently produces a prediction, with a 0% no-prediction rate, as it is trained on grounding datasets that enforce object localization for every referred object. As a result, CogVLM [35] does not fail to detect objects but may still struggle with accurate label assignment. Additionally, LISA [13] exhibits the highest value of $p(x_{n \rightarrow bg} | x_{k \rightarrow k})$, suggesting weaker object localization capabilities compared to other models.

5.3. VLMs Leverage Spatial Context To Reduce Ambiguity

We examine the models' ability to leverage spatial context to disambiguate between multiple novel objects when more than one is present in the scene. Table 2 illustrates model performance in the 1K-2U setting, where two novel objects are present in the scene. When scene context is absent, ambiguity arises between the two novel objects, leading most models to exhibit a high ambiguity score which is an expected outcome, as there are no additional cues available for distinguishing between the two novel objects. However, when scene context is provided, the ambiguity score drops significantly across all models, demonstrating that models can effectively leverage spatial information to resolve ambiguity and correctly assign novel labels.

Gemini [32] benefits the most from the additional spatial input, achieving the highest spatial reasoning score, indicating strong contextual refinement in label assignment. Other models also show improvements in spatial reasoning, though to varying degrees. CogVLM [35], once again, ranks the lowest in its ability to effectively utilize spatial context for reasoning.

5.4. Understanding The Use of Object Names in the ME Task

In this subsection, we analyze additional factors that may influence model performance, beyond solely ME bias and spatial reasoning, using the 2K-1U setting. We select this setting because it presents a greater visual challenge compared to 1K-1U, providing a larger margin for improvement and deeper insights into model behavior. Additionally, this setting is inherently deterministic: a model leveraging ME bias can solve the task without relying on extra contextual information from the scene. This allows us to systematically evaluate the impact of additional information and determine whether it contributes to performance improvements.

In Figure 6, we present the models' performance across three different prompt settings: (1) Question-Only Prompt: The prompt consists solely of the query, e.g., "Where is the dax?"; (2) Minimal Scene Context: The prompt includes only the object names present in the scene, e.g., "There are three objects in the scene: a dog, a cat, and a dax. Where is the dax?"; (3) Full Scene Description: The models receive a detailed scene description as outlined in Section 3.3.

Table 2. **Impact of Spatial Scene Input on Ambiguity & Spatial Reasoning Scores.** We compare settings with and without spatial scene input, evaluating models based on two key metrics: Ambiguity Score, which measures the ability to disambiguate novel objects (lower is better), and Spatial Reasoning Score, which reflects how effectively models utilize spatial context to improve performance, with respect to the setting without spatial context (higher is better). Bold values indicate the most desirable outcomes.

Method	CogVLM	Sa2VA	OMG-LLaVA	Gemini	LISA	F-LMM
Ambiguity without scene dec. ↓	0.525	0.480	0.500	0.510	0.394	0.510
Ambiguity with scene dec. ↓	0.384	**0.258**	0.345	**0.257**	0.283	0.365
Spatial Reasoning (Relative to w/o scene dec.) ↑	0.314	0.914	1.012	**11.167**	1.853	1.314
Spatial Reasoning (Absolute) ↑	0.117	0.203	0.165	**0.268**	0.265	0.230

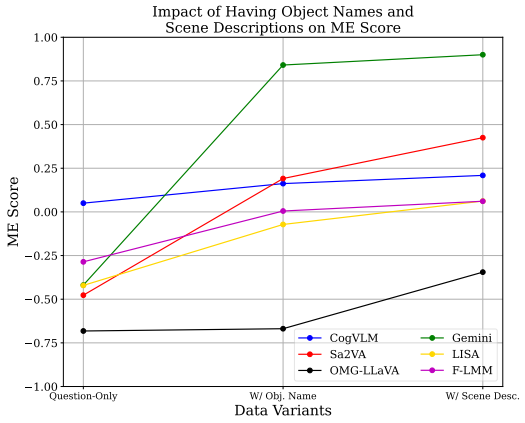


Figure 6. **Impact of Having Object Names and Scene Descriptions on ME Score** comparing between Question-Only language prompt (e.g. Where is the dax?); Minimal Scene Context (e.g. There are three objects in the scene: dog, a cat, and a dax. Where is the dax?) and Full Scene Description: The models receive a detailed scene description

By comparing performance between (1) Question-Only and (2) Minimal Scene Context, we assess the impact of knowing the object names present in the scene. The difference between (2) Minimal Scene Context and (3) Full Scene Description further evaluates the models' ability to leverage spatial reasoning when additional context is provided.

All models show performance improvements when provided with Minimal Scene Context (i.e., object names). This result aligns with expectations: while the additional information consists solely of object labels, it implicitly signals to the model that "dax" is distinct from "dog" and "cat", reinforcing the mutual exclusivity assumption and aiding in accurate object-label associations.

With the exception of CogVLM [35], which performs similarly to the Minimal Scene Context setting, all other models show improved performance in the Full Scene Description case, where spatial information is provided. This demonstrates the benefit of incorporating spatial context in aiding object reasoning and disambiguation.

6. Limitations

Multiple additional factors influence mutual exclusivity bias in real-world settings, including social interaction and temporal ambiguity, which we leave to future work. Although our setup captures uncertainty reasoning, scene-level spatial descriptions represent only one possible way to resolve ambiguity. In practice, children and adults rely on many other mechanisms, such as accumulated experience over time and broader reasoning strategies. Future work could investigate alternative ways of resolving ambiguity under uncertainty. As with all studies on LLMs and VLMs, performance can also be highly sensitive to prompt design, and different models may respond best to different prompts. In this work, we use the prompts recommended by the original authors for each method in object detection and segmentation settings. Future research could explore alternative prompting strategies both to better understand model behavior and to further improve their reasoning performance.

7. Conclusion

We introduce MEBench, a novel benchmark designed to study mutual exclusivity (ME) bias and to extend this analysis toward spatial reasoning in object-label mapping tasks for SOTA VLMs. We assess several SOTA VLMs on MEBench using a scalable synthetic data generation pipeline. These models exhibit weak ME bias, and their performance declines sharply even under modest increases in scene complexity. Furthermore, providing additional scene context helps reduce ambiguity and improves performance. However, these models remain far from an ideal solution, one that would seamlessly integrate strong object localization, ME bias, and spatial reasoning. Bridging this gap remains a crucial direction for future research, as these properties are essential for developing more advanced, human-like reasoning in AI models. We hope that MEBench will encourage the AI community to systematically explore the interplay between ME bias, spatial reasoning, and multimodal learning, and to develop models with stronger cognitive capabilities.

References

[1] Blender, https://blender.org/. 2, 3, 4

[2] Geoshapes add-on, https://blendermarket.com/products/geoshapes-parametric-geometry-node-objects. 4

[3] Harsh Agrawal, Eli A. Meirom, Yuval Atzmon, Shie Mannor, and Gal Chechik. Known unknowns: Learning novel concepts using reasoning-by-elimination. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 504–514. PMLR, 2021. 2

[4] Anthropic. Claude 3. https://www.anthropic.com/index/claude, 2024. Accessed: 2024-02-28. 3

[5] Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund, Daniel LK Yamins, James J DiCarlo, Josh McDermott, Antonio Torralba, et al. The threedworld transport challenge: A visually guided task-and-motion planning benchmark for physically realistic embodied ai. *arXiv preprint arXiv:2103.14025*, 2021. 2

[6] Kanishk Gandhi and Brenden M Lake. Mutual exclusivity as a challenge for deep neural networks. *Advances in Neural Information Processing Systems*, 33:14182–14192, 2020. 1, 2, 5

[7] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. 2

[8] Felix Hill, Olivier Tieleman, Tamara Von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. Grounded language learning fast and slow. *arXiv preprint arXiv:2009.01719*, 2020. 2

[9] Jessica S Horst and Michael C Hout. The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior research methods*, 48(4):1393–1409, 2016. 3

[10] Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. Mewl: Few-shot multimodal word learning with referential uncertainty. *arXiv preprint arXiv:2306.00503*, 2023. 2

[11] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19056–19065, 2022. 3

[12] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021. 2

[13] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2, 5, 6, 7

[14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 5

[15] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 3

[16] Ellen M Markman and Gwyn F Wachtel. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2):121–157, 1988. 2

[17] Ellen M Markman, Judith L Wasow, and Mikkel B Hansen. Use of the mutual exclusivity assumption by young word learners. *Cognitive psychology*, 47(3):241–275, 2003. 2

[18] Bob McMurray. Defusing the childhood vocabulary explosion. *Science*, 317(5838):631–631, 2007. 1

[19] OpenAI. Gpt-4 with vision (gpt-4v). https://openai.com/research/gpt-4, 2023. Accessed: 2024-02-28. 3

[20] Deepan Chakravarthi Padmanabhan, Shruthi Gowda, Elahe Arani, and Bahram Zonooz. Lsfsl: Leveraging shape information in few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4970–4979, 2023. 2

[21] Xinyu Pi, Mingyuan Wu, Jize Jiang, Haozhen Zheng, Beitong Tian, Chengxiang Zhai, Klara Nahrstedt, and Zhiting Hu. Uouo: Uncontextualized uncommon objects for measuring knowledge horizons of vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6432–6441, 2024. 3

[22] Shitala Prasad, Yiqun Li, Dongyun Lin, and Aiyuan Guo. Implicit shape biased few-shot learning for 3d object generalization. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3436–3440. IEEE, 2022. 2

[23] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21783–21794, 2024. 3

[24] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2, 3

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language

understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[28] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 2

[29] Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3): 1558–1568, 2008. 2

[30] Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1798–1808, 2021. 2, 3

[31] Stefan Stojanov, Anh Thai, Zixuan Huang, and James M. Rehg. Learning dense object descriptors from multiple views for low-shot category generalization. In *Advances in Neural Information Processing Systems*, pages 12566–12580, 2022. 3

[32] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. 3, 5, 7

[33] Anh Thai, Ahmad Humayun, Stefan Stojanov, Zixuan Huang, Bikram Boote, and James M Rehg. Low-shot object learning with mutual exclusivity bias. *Advances in Neural Information Processing Systems*, 36:70208–70228, 2023. 2

[34] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2025. 5

[35] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for large language models. 2023. 3, 5, 6, 7, 8

[36] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav ARORA, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jerome Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *Advances in Neural Information Processing Systems*. 3

[37] Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. *arXiv preprint arXiv:2406.05821*, 2024. 3, 5, 6

[38] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint arXiv: 2303.04803*, 2023. 2

[39] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 2, 3, 5

[40] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37:71737–71767, 2025. 3, 5

[41] Qingnan Zhou and Alec Jacobson. Thingi10k: A dataset of 10,000 3d-printing models. *arXiv preprint arXiv:1605.04797*, 2016. 4