CounterBench: A Controllable Counterfactual Testbed Reveals Systematic Reasoning Failures in Vision-Language Models

Aayam Bansal Ishaan Gangwani
Synthetic Sciences
{aayam, ishaan}@syntheticsciences.ai

Abstract

Vision-language models (VLMs) achieve impressive accuracy on standard visual question answering benchmarks, yet it remains unclear whether they reason about scenes or merely pattern-match from surface cues. We introduce CounterBench, a fully synthetic, controllable benchmark for evaluating counterfactual consistency *in VLMs. For each of 550 test items, we programmatically generate a paired scene: an original and an* intervened *variant where exactly one semantic property is changed (an object's position, color, count, containment, or a causal link). We pose the same question to both images and measure whether the model's answer changes* if and only if *the intervention is relevant—the* Counterfactual Consistency Score *(CCS). Evaluating four state-of-the-art VLMs (GPT-4o, Gemini 2.0 Flash, Qwen2.5-VL-72B, GPT-4o mini), we find that while all models achieve 85–91% single-image accuracy, CCS drops to 80–90%, with dramatic category-specific failures:* all *models score* $\leq$*66% on causal arrow-following tasks, GPT-4o mini achieves only 48% CCS on counting, and even the strongest model (Qwen2.5-VL) reaches only 89.5% overall CCS. Crucially, spatial and containment reasoning are near-perfect (97–100%), revealing that failures are* selective*, not uniform. Our results demonstrate that a cheap, fully controllable testbed—generated in minutes with standard Python libraries—can surface systematic VLM failures invisible to standard benchmarks. We release the full benchmark, generation code, and evaluation pipeline.*

1. Introduction

The rapid progress of vision-language models (VLMs) such as GPT-4o [15], Gemini [5], and Qwen2.5-VL [2] has yielded systems that can describe complex images, answer nuanced questions, and reason about visual scenes. Yet a growing body of evidence suggests that strong benchmark performance may mask fundamental reasoning deficiencies. Winoground [18] revealed near-chance compositional reasoning; the ARO benchmark [22] showed VLMs behave as "bags of words"; BLINK [3] demonstrated that even GPT-4V performs only slightly above chance on core perceptual tasks; and MMVP [19] exposed "CLIP-blind" patterns that propagate through entire model families.

A key limitation of existing benchmarks is that they test whether models give *correct answers*, but not whether models *reason correctly*. A model might answer "the ball is left of the cup" correctly by exploiting statistical regularities in its training data, without genuinely parsing the spatial relationship. To distinguish true reasoning from pattern matching, we need *counterfactual* evaluation: if we intervene on the scene—say, swap the positions of ball and cup—does the model's answer flip accordingly? And equally important: if we make an *irrelevant* change (e.g., alter the background color), does the answer *stay the same*?

This counterfactual framing, grounded in Pearl's interventionist theory of causation [16] and developmental studies of causal cognition [6, 17], provides a principled test of whether models maintain internally consistent world models. Counterfactual consistency has been studied in NLP [14] and for text-image matching [20, 23], but no existing benchmark provides a fully controllable, programmatically generated testbed with paired visual interventions and a formal consistency metric.

We introduce CounterBench, a benchmark with the following properties:

1. **Paired counterfactual design.** Each item consists of an original image and an intervened image where exactly one semantic property is changed, plus a question whose correct answer is deterministic for both images.
2. **Five cognitive dimensions.** We test spatial relations, attribute binding, counting, containment, and causal reasoning (arrow-following, occlusion, compositionality).
3. **Negative controls.** 50 items where the intervention is *irrelevant* (adding a distant distractor), testing whether models maintain stability.
4. **Counterfactual Consistency Score (CCS).** The percentage of pairs where the model answers *both* the original and intervened image correctly—capturing not just

accuracy but *reasoning consistency*.

5. **Zero cost, full control.** All 1,100 images are generated in <1 minute using PIL (Python Imaging Library), with deterministic ground truth.

Our evaluation of four VLMs reveals selective, systematic failures:

- All models achieve near-perfect CCS on spatial (100%) and containment (97–100%) tasks—these are "solved."
- Causal reasoning (arrow-following) is universally weak: every model scores ≤66% CCS.
- Counting consistency varies dramatically: Qwen2.5-VL achieves 97% but GPT-4o mini only 48%.
- Attribute binding shows unexpected model-specific failures: Gemini 2.0 Flash drops to 79% despite strong overall accuracy.
- The *consistency gap*—the difference between single-image accuracy and CCS—ranges from 0 to 14 percentage points, revealing that some models give correct answers to individual images but inconsistent answers across the pair.

These findings demonstrate that a cheap, fully controllable testbed—requiring no training data, no crowdsourcing, and no GPU compute—can reveal systematic VLM failures that are invisible to standard benchmarks.

2. Related Work

Compositional reasoning benchmarks. A rich line of work probes VLM compositionality. Winoground [18] uses 400 human-curated image-caption pairs requiring word-order sensitivity, finding VLMs near chance. CREPE [13] tests systematicity and productivity of compositionality across 370K+ pairs. SugarCrepe [7] showed that prior benchmarks were "hackable"—text-only models could outperform VLMs—and proposed adversarial refinement. The ARO benchmark [22] demonstrated bag-of-words behavior with 50K+ test cases. ConMe [8] uses VLM-to-VLM adversarial conversations to generate harder compositional challenges. These benchmarks use *natural* images with text-based perturbations; by contrast, CounterBench uses *fully synthetic* images with *visual* interventions, providing complete control over the counterfactual manipulation.

Controlled and synthetic evaluation. CLEVR [9] pioneered programmatic scene generation for visual reasoning, though it predates modern VLMs. VisMin [1] generates minimal-change image pairs using diffusion models. CounterCurate [23] uses GLIGEN and DALL-E 3 for counterfactual image generation. VSR [12] tests spatial reasoning with natural images. Our work differs from diffusion-based approaches in that our scenes are *deterministically generated* with exact ground truth—no generation artifacts, no ambiguity, and trivial to reproduce.

Counterfactual reasoning. Counterfactual VQA [14] applies causal interventions to reduce language bias in VQA. EqBen [20] tests equivariant similarity—whether VLM similarity scores change faithfully under semantic modifications. NaturalBench [10] pairs questions with two images yielding different answers. Our CCS metric most closely relates to EqBen's equivariance concept, but extends it from similarity scoring to open-ended question answering across multiple cognitive dimensions.

Cognitive evaluation of VLMs. BLINK [3] reformulates classic CV tasks as VQA, finding GPT-4V at 51% vs. human 96%. MMMU [21] tests expert-level reasoning across 30 subjects. POPE [11] measures object hallucination. Gavrikov *et al*. [4] study texture-shape bias inheritance in VLMs. Our work contributes a cognitively-grounded evaluation that probes *counterfactual* reasoning—a core capacity identified in developmental psychology [6]—using a formally precise, zero-cost methodology.

3. The CounterBench Benchmark

3.1. Design Principles

CounterBench is built on four principles: (1) **counterfactual pairing**: every test item is a paired $(I_{\text{orig}}, I_{\text{int}})$ with a shared question q; (2) **deterministic ground truth**: correct answers are computed from the generation parameters, eliminating annotation noise; (3) **minimal intervention**: each intervention changes exactly one semantic property; (4) **full reproducibility**: the entire benchmark is generated by a single Python script with a fixed random seed.

3.2. Scene Generation

All scenes are 512×512 pixel images with a light gray background, containing colored geometric shapes (circles, squares, triangles, diamonds) in 8 distinct colors. We use PIL (Python Imaging Library) to render shapes, containers (rectangles), arrows, and text labels. The generation script produces 1,100 images (550 original-intervened pairs) in under 60 seconds on a single CPU.

3.3. Task Categories

We define five task categories (100 pairs each) plus 50 negative control pairs:

Spatial Relations (100 pairs). Two labeled shapes (A and B) are placed in a spatial arrangement. Question: "Is A to the {left/right/above/below} of B?" Intervention: swap the positions of A and B. Expected: answer flips from "yes" to "no."

Attribute Binding (100 pairs). Multiple shapes of different colors are arranged on the canvas. Question: "What

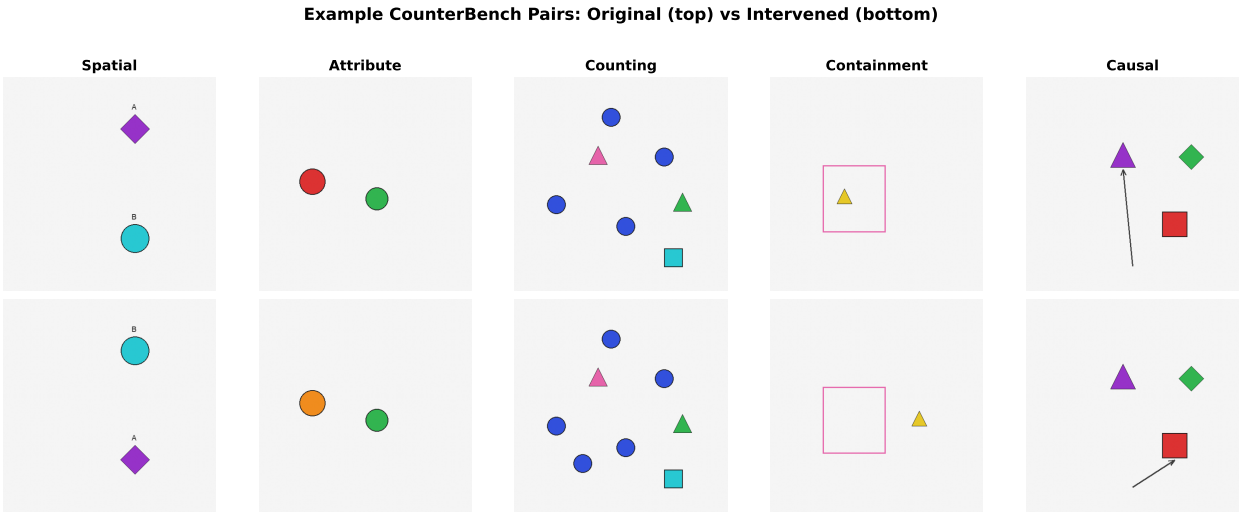

Figure 1. **Example CounterBench pairs** across five task categories. **Top**: original scenes. **Bottom**: intervened scenes with exactly one semantic property changed. Questions and ground-truth answers are deterministic for each image.

color is the {position} {shape}?" Intervention: change the target shape's color. Expected: answer changes to the new color.

Counting (100 pairs). A target set of identical shapes and a distractor set. Question: "How many {color} {shape}s are there?" Intervention: add or remove one target shape. Expected: count changes by ± 1.

Containment (100 pairs). A shape and a rectangular container. Question: "Is the {shape} inside the container?" Intervention: move the shape inside $\leftrightarrow$ outside. Expected: answer flips.

Causal/Compositional (100 pairs). Three sub-tasks: (a) *Arrow pointing* (34 pairs): an arrow points to one of three shapes; intervention redirects the arrow; question asks which shape is pointed to. (b) *Occlusion* (33 pairs): two overlapping shapes; intervention removes the front shape; question asks how many shapes are visible. (c) *Compositional* (33 pairs): two shapes in a vertical arrangement; intervention swaps vertical positions; question asks about the spatial composition.

Negative Controls (50 pairs). A single shape in the center. Question: "What color is the {shape}?" Intervention: add an unrelated small shape in a corner (far from the target). Expected: answer should *not* change.

3.4. Metrics

We define three metrics for each model M on a set of paired items $\mathcal{P}$:

Original accuracy: $\text{Acc}_{\text{orig}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}[M(I_p^{\text{orig}}, q_p) = a_p^{\text{orig}}]$

Intervened accuracy: $\text{Acc}_{\text{int}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}[M(I_p^{\text{int}}, q_p) = a_p^{\text{int}}]$

Counterfactual Consistency Score:

$$\text{CCS} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{1}\big[M(I_p^{\text{orig}}, q_p) = a_p^{\text{orig}}\big] \cdot \mathbf{1}\big[M(I_p^{\text{int}}, q_p) = a_p^{\text{int}}\big] \tag{1}$$

CCS requires the model to be correct on *both* the original and intervened image. This is strictly harder than either single-image accuracy alone, and captures whether the model's internal representation is sensitive to the specific intervention performed. For negative controls, the intervened ground truth equals the original, so CCS measures answer stability under irrelevant changes.

4. Experiments

4.1. Setup

Models. We evaluate four VLMs via the OpenRouter API: (1) **GPT-4o** [15], OpenAI's flagship multimodal model; (2) **Gemini 2.0 Flash** [5], Google's efficient multimodal model; (3) **Qwen2.5-VL-72B** [2], Alibaba's open-weight VLM; (4) **GPT-4o mini**, a smaller variant of GPT-4o. All models are queried with temperature 0 and instructed to give short, direct answers.

Answer normalization. We normalize responses by extracting yes/no keywords, parsing numbers (including word-to-digit), and matching color names. For causal-arrow tasks, we check whether both the color and shape are mentioned.

Table 1. **Overall CounterBench results.** Orig/Int = accuracy on original/intervened images. Both = correct on both. CCS = Counterfactual Consistency Score (Eq. 1).

Model	Orig (%)	Int (%)	Both (%)	**CCS** (%)
GPT-4o	88.5	88.7	86.0	86.0
Gemini 2.0 Flash	90.4	90.0	87.8	87.8
Qwen2.5-VL-72B	**91.4**	**90.6**	**89.5**	**89.5**
GPT-4o mini	84.5	84.7	79.6	79.6

Compute. The entire evaluation—generating all images and running all 4,400 model queries (4 models $\times$ 1,100 images)—requires no GPU and completes via API in approximately 2 hours. Total API cost: <$12.

4.2. Main Results

Table 1 shows overall results. Qwen2.5-VL-72B achieves the highest CCS (89.5%), followed by Gemini 2.0 Flash (87.8%), GPT-4o (86.0%), and GPT-4o mini (79.6%). Notably, the gap between original accuracy and CCS is small overall (1–5 pp), suggesting that most errors are consistent across the pair rather than arising from inconsistency. However, this aggregate masks dramatic category-level variation, as we show next.

4.3. Category-Level Analysis

Table 2 reveals the most striking finding of our work: **VLM failures are selective, not uniform.** We identify three tiers:

Tier 1: Solved (CCS $\geq$92%). Spatial relations and containment are near-perfect across all models. Even GPT-4o mini achieves 92% on spatial and 97% on containment. These tasks require basic visual parsing of labeled shapes and rectangular boundaries—capacities that current VLMs handle reliably.

Tier 2: Model-dependent (48–97%). Counting and attribute binding show dramatic inter-model variation. For counting, Qwen2.5-VL-72B achieves 97% CCS while GPT-4o mini manages only 48%—a **49 percentage point gap**. For attribute binding, GPT-4o leads at 94% while Gemini 2.0 Flash drops to 79%, despite Gemini's higher overall accuracy. This suggests that different architectures and training recipes produce distinct cognitive profiles.

Tier 3: Universally hard (CCS $\leq$66%). Causal reasoning—primarily arrow-following—is the universal failure mode. Every model, regardless of size or architecture, scores 64–66% CCS. This means roughly one-third of the time, models either follow the wrong arrow or fail to update their answer when the arrow is redirected. This finding is consistent with prior work showing VLMs struggle with directional and relational reasoning [3, 12].

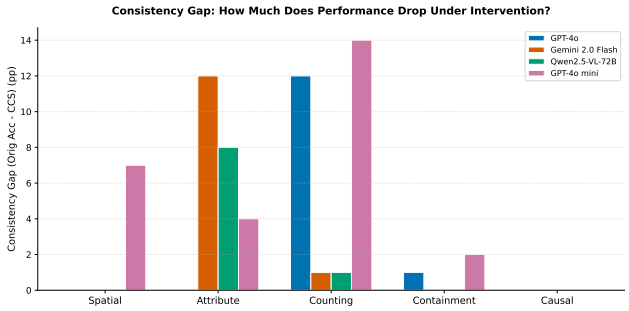

Figure 2. **Consistency gap** (original accuracy − CCS) by category. Larger gaps indicate models that give correct answers to individual images but fail to maintain consistency across the paired intervention.

4.4. The Consistency Gap

Figure 2 shows the consistency gap per category. Key observations:

- **Counting** has the largest gap for GPT-4o (12 pp) and GPT-4o mini (14 pp): these models often get the original count right but fail on the intervened count (or vice versa).
- **Attribute binding** shows a notable gap for Gemini (12 pp), suggesting Gemini sometimes recognizes colors correctly but is confused when the same shape appears in a different color.
- **Spatial and containment** have near-zero gaps across all models—when these models succeed on one image, they consistently succeed on both.

4.5. Negative Control Analysis

Figure 3 shows that all models maintain high stability (90–96%) on negative controls, where an unrelated shape is added far from the target. This confirms that the failures observed in Table 2 are genuine reasoning failures, not artifacts of visual distraction sensitivity.

4.6. Qualitative Error Analysis

Figure 4 visualizes the CCS profile as a radar chart. All models exhibit a characteristic "dent" in the causal dimension. GPT-4o mini shows an additional dent in counting, while Gemini shows one in attribute binding. Qwen2.5-VL-72B has the most uniform profile, but still shares the causal weakness.

Manual inspection of causal-arrow errors reveals two failure patterns: (1) **arrow blindness**: the model names a shape but ignores the arrow entirely, defaulting to the most salient or centered object; (2) **sticky answers**: the model gives the same answer for both original and intervened im-

Table 2. **CCS by task category (%).** Green : ≥90%. Yellow : 70–89%. Red : <70%. All models fail on causal tasks (≤66%). Counting reveals the largest inter-model variance (48–97%).

Model	Spatial	Attribute	Counting	Containment	Causal	Neg. Control
GPT-4o	100.0	94.0	67.0	99.0	66.0	94.0
Gemini 2.0 Flash	100.0	79.0	93.0	100.0	66.0	90.0
Qwen2.5-VL-72B	100.0	84.0	97.0	100.0	66.0	92.0
GPT-4o mini	92.0	89.0	48.0	97.0	64.0	96.0

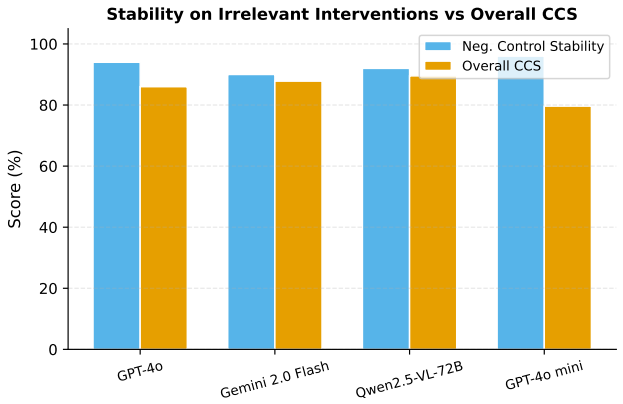

Figure 3. **Negative control stability.** Models should maintain the same answer when irrelevant distractors are added. All models achieve 90–96% stability, indicating they rarely hallucinate or flip answers due to irrelevant visual changes.

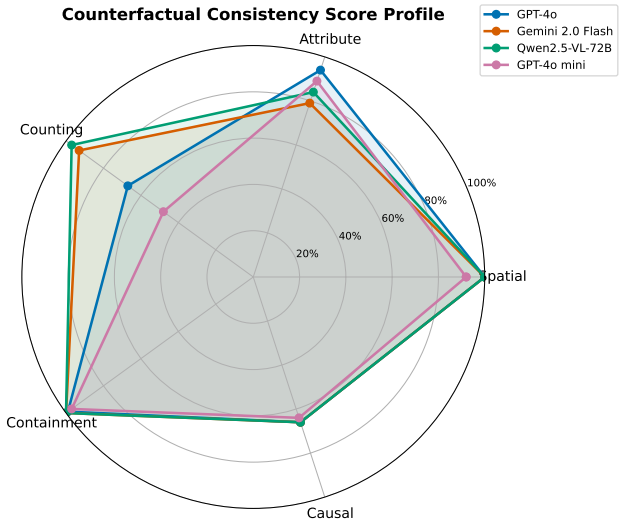

Figure 4. **CCS profile** across five task categories. The "dent" in the causal dimension is universal; other dimensions show model-specific strengths.

ages, suggesting it encodes the scene layout but not the arrow direction.

For counting errors, models often err by ±1 in scenes with 4+ objects, consistent with known subitizing limits [3] and the difficulty of precise enumeration in VLMs.

5. Discussion

Why counterfactual consistency matters. Standard accuracy metrics credit models for getting single questions right, but do not test whether the model's understanding is coherent across related scenarios. A model that answers "3 circles" in the original and "3 circles" in the intervened image (where there are now 4) scores 50% accuracy but 0% CCS—revealing that it pattern-matches rather than counts. Our CCS metric penalizes such inconsistency, providing a stricter test of genuine visual reasoning.

The causal reasoning gap. The universal ≤66% CCS on causal tasks is our most striking finding. Arrow-following requires the model to: (1) detect the arrow, (2) trace its direction, (3) identify the target. This is precisely the kind of relational, directional reasoning that BLINK [3] and VSR [12] found lacking. Our results show this deficit persists even in 2025-era models, suggesting it may be architectural rather than a matter of scale.

Cheap benchmarking as a methodology. CounterBench demonstrates that significant scientific insights can be obtained from trivially simple stimuli. The entire benchmark—generation, inference, and analysis—costs <$12 and requires no GPU. This makes counterfactual testing accessible to any researcher and suggests a methodology for rapid, targeted probing of VLM capabilities as new models emerge.

Implications for cognitive foundations. From a cognitive science perspective [17], our results suggest VLMs have acquired "core knowledge" of spatial relations and containment but lack robust causal and quantitative reasoning. This mirrors developmental trajectories where children master spatial before causal reasoning [6], though the analogy should not be taken too literally.

Limitations. (1) Our scenes are simple geometric shapes, not natural images. While this is by design (for full control),

it means our findings may not directly transfer to natural scenes. (2) We test four models; broader evaluation across open-source VLMs would strengthen the findings. (3) The causal subtask (arrow-following) may conflate arrow detection with causal reasoning proper. (4) Answer normalization introduces potential parsing errors, though our conservative matching reduces false positives.

6. Conclusion

We introduced CounterBench, a fully synthetic, zero-cost benchmark for evaluating counterfactual consistency in vision-language models. By generating 550 paired images with precise visual interventions across five cognitive dimensions, we measure whether VLMs maintain consistent reasoning when scenes are modified. Our evaluation of four state-of-the-art VLMs reveals a heterogeneous failure landscape: spatial and containment reasoning are near-perfect, counting and attribute binding are model-dependent, and causal reasoning is universally weak. The Counterfactual Consistency Score (CCS) provides a complementary metric to standard accuracy that captures reasoning coherence, and the $\leq$66% causal CCS shared by all models points to a systematic architectural limitation. We release all code, images, and evaluation scripts, and argue that cheap, controllable testbeds are an underutilized tool for understanding the cognitive foundations of multimodal AI.

References

[1] Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual minimal-change understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[2] Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

[4] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, et al. Can we talk models into seeing the world differently? *International Conference on Learning Representations (ICLR)*, 2025.

[5] Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2024.

[6] Alison Gopnik, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review*, 111(1):3–32, 2004.

[7] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[8] Irene Huang, Wei Lin, M. Jehanzeb Mirza, et al. ConMe: Rethinking evaluation of compositional reasoning for modern VLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[9] Justin Johnson, Bharath Hariharan, Laurence van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[11] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[12] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics (TACL)*, 2023.

[13] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[14] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] OpenAI. GPT-4 technical report. In *arXiv preprint arXiv:2303.08774*, 2023.

[16] Judea Pearl. Causality: Models, reasoning, and inference. 2009.

[17] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

[18] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[21] Xiang Yue et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert

AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[22] 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? In *International Conference on Learning Representations (ICLR)*, 2023.

[23] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. CounterCurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.