COUNTERCOUNT: DIAGNOSING KNOWLEDGE BIAS IN HOW VISION-LANGUAGE MODELS COUNT ATTRIBUTES

Anonymous authors
Paper under double-blind review

ABSTRACT

Do Vision-Language Models (VLMs) count what they see, or what they know about the world? We introduce *CounterCount*, a diagnostic dataset of paired factual and counterfactual (CF) images, in which the CF image contains an atypical count of a subject's attribute (*e.g.*, a rabbit with four ears), together with annotations localizing the counted attributes. Humans answer both types of images nearly perfectly (98.9% and 97.5%), but across 16 VLMs (14 open-weight and 2 closed-source), average accuracy drops from 91.6% on factual to 49.5% on CF images, with most errors stemming from responses that match the canonical count, therefore revealing a bias towards knowledge priors rather than reliance on visual information. Based on these findings, we perform an analysis that reveals two interesting insights behind this gap. First, models exhibit a "System-1"-like response behavior, falling back on canonical counts after a shallow recognition of the subject. Second, model activations partially contain the visual evidence necessary to answer correctly, yet models underutilize it, attending less to the counted attributes when they provide biased answers. Input manipulations that break visual context, introduce cueing or reasoning, or direct amplification of the model's internal attention to relevant regions can reduce such biases. *CounterCount* goes beyond showing that VLMs fall back on knowledge priors, revealing that once a familiar context is recognized, models favor prior knowledge over available visual evidence rather than failing to encode that evidence.

1 INTRODUCTION

Vision-Language Models (VLMs) have advanced rapidly in recent years (Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023; Bai et al., 2023; 2025; Kamath et al., 2025; Wang et al., 2025), achieving strong performance on Visual Question Answering (VQA) and multimodal reasoning (Liu et al., 2023; Zhang et al., 2024; Yue et al., 2024; Liu et al., 2024b; Antol et al., 2015; Hudson & Manning, 2019; Lu et al., 2024; Yu et al., 2024). Yet strong benchmark performance can mask a reliance on knowledge priors, learned from the statistical regularities of large-scale text corpora and the visual patterns encountered during training (Vo et al., 2025c; Sheng et al., 2019; Gallegos et al., 2024; Vo et al., 2025a;b; Rahmanzadehgervi et al., 2024; Lee et al., 2025; Fu et al., 2025), rather than genuine grounding in the image.

In general VQA, such reliance is often easy to expose: if a model describes the Statue of Liberty as holding a torch when the image shows a sword, the prior is clearly at work (Lee et al., 2025). Counting attributes is more subtle, as a wrong count may stem either from a default to the subject's canonical count or from a genuine limitation in counting ability. Studies on general object counting, where no canonical prior is involved, help separate the two. While some VLMs can still make errors at small counts (Anh et al., 2026; Alghisi et al., 2025), these studies consistently find that counting accuracy degrades as the number of objects increases (Sengupta et al., 2025; Qharabagh et al., 2025; Alghisi et al., 2025). Small counts therefore offer the setting where counting ability is least likely to be the bottleneck, and where errors most directly reflect prior-driven defaults. But the few datasets that explicitly study counting bias (Chen et al., 2026; Vo et al., 2025b) are limited in scale, often involve large counts that confound bias with counting difficulty, or contain ambiguous images.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

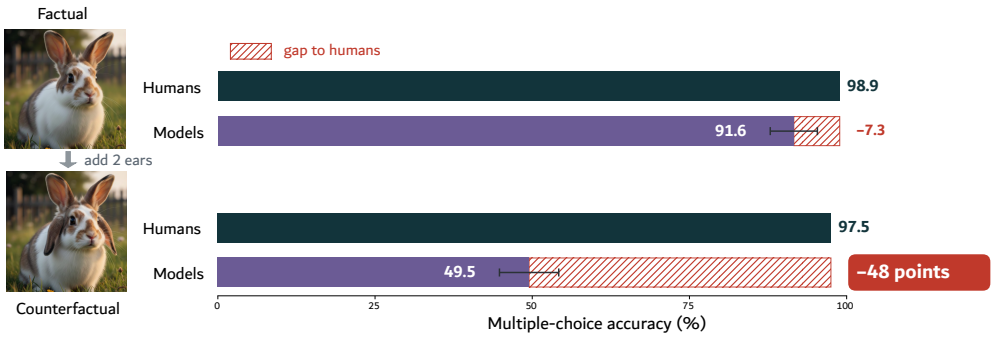


Figure 1: VLMs nearly match humans on factual images but collapse on counterfactual ones, while humans remain near ceiling, suggesting a reliance on knowledge priors over visual evidence.

To this end, we introduce *CounterCount* (Section 2), a diagnostic dataset of paired factual and counterfactual (CF) images, where each pair targets an attribute with a strong canonical prior (*e.g.* ears, wheels, or legs) at a small count. A controlled human study confirms that the task is trivial: humans achieve near-perfect accuracy on both factual and CF images (98.9% and 97.5%; Figure 1), showing that CF images pose no additional perceptual difficulty. On the other hand, VLMs approach human accuracy on factual images (91.6%), where the canonical count is also the correct answer, they collapse to 49.5% on CF images, with over *three quarters* of their CF errors returning exactly the canonical count.

To understand this failure, we conduct a systematic analysis that reveals two behaviors behind it (Section 4). **First, models exhibit System-1-like behavior** (Section 4.1): the bias is triggered by a shallow recognition of the subject in the image rather than by how it is named in the prompt, and disturbing this recognition shortcut through blurring or cueing, or encouraging System-2-like reasoning (Kahneman, 2011) through thinking mode, moves predictions away from the canonical count. **Second, models have the visual evidence but underuse it** (Section 4.2): failures coincide with under-attention to the attributes they are asked to count, and redirecting attention toward them helps models act on evidence they already have. These findings suggest that recognizing the subject leads models to recall rather than count, even though the information required to count correctly is available to them.

In summary, our main contributions are:

- **CounterCount.** A diagnostic dataset of 554 manually verified factual/CF image pairs across 379 subjects, 228 attribute types, and 17 categories, with localized oracle annotations (masks and bounding boxes) for fine-grained analysis of counting bias (Section 2).
- **Human and model evaluation.** An evaluation of humans and 16 VLMs (14 open-weight, 2 closed-source), showing that humans achieve near-ceiling accuracy on both factual and CF images while VLMs drop sharply on CF images, demonstrating that model failures are not due to task difficulty (Figure 1).
- **Diagnostic analysis.** A systematic exploration yielding two key findings: (1) models exhibit System-1-like behavior, falling back on canonical counts based on knowledge after a shallow recognition of the subject, and (2) they overlook available visual evidence, attending less to count-relevant regions when they fail (Section 4).

2 CounterCount Dataset

To test whether VLMs ground counting predictions in visual evidence or rely on canonical object priors, we construct paired factual and CF images. An image pair differs only in a localized count-relevant edit, preserving viewpoint, background, lighting, and global composition while altering the number of a structural attribute (*e.g.* ears, legs, wheels). The edited regions are fully visible and manually verified, so the target count is explicit and unambiguous. This minimal design helps attribute performance differences between factual and CF images to conflicts between visual evidence



Figure 2: **CounterCount dataset construction and evaluation pipeline.** The pipeline follows four steps: (1) **Factual Image Generation**, where an instruction generates an image reflecting a canonical attribute, such as a cat with one tail; (2) **CF Image Generation**, where a numerical modifier (e.g., +2 tails) is applied to deliberately alter the canonical attribute and produce an edited counterpart; (3) **Human Verification**, where the image pair is reviewed to confirm the counts, resulting in either a pass or a command to regenerate; (4) **Question Formats**, Open-Ended (OE), and Multiple Choice (MCQ)

and object-level priors rather than incidental scene variation. To support analysis beyond answer accuracy, *CounterCount* also provides localized masks and bounding boxes (BBs) for the modified count-relevant regions, enabling region-level analysis and controlled diagnosis during inference.

2.1 Dataset Construction

We chose common, widely recognized subjects to ensure broad coverage across familiar attributes, starting from broad groups (*e.g.*, animals) and expanding into finer categories (*e.g.*, mammals, birds). *CounterCount* spans seventeen semantic categories across biological, manufactured, and functional domains, containing 379 unique subjects (*e.g.*, dog, airplane, Starbucks logo) and 228 countable attribute types (e.g., leg, wing, crown), capturing a diverse set of canonical object priors. The full category list and per-category breakdown are provided in Appendix A.1.

Attribute Selection and Canonical Priors. For each subject, we select attributes with widely recognized canonical counts and edit them to create counts that directly conflict with the expected object structure, such as a giraffe with five heads or a car with six wheels. This places visual evidence in conflict with learned object priors, so that edited instances cannot be answered from prior knowledge alone; instead, models must ground their predictions in the visible count-relevant evidence.

Image Generation and Quality Control. We generate a photorealistic factual image with canonical counts using Nano-Banana [1]. Then, we prompt the same model for a CF variant that alters one count-relevant attribute while preserving viewpoint, lighting, background, and object identity. Each image pair is manually verified for photorealistic quality, semantic coherence, and visibility of the edited attribute; pairs that fail verification are regenerated until all criteria are met. This yields 554 CF images paired with their factual counterparts (1,003 unique images in total since a single factual image can serve as the base for multiple CF edits targeting different attributes), all at a standardized 480×480 resolution. Further pipeline details and examples are provided in Figure 2 and Appendix A.3.

Question Format. We evaluate each image pair by asking models how many instances of the target attribute the subject has. All questions are generated from fixed templates filled with each image's manually annotated metadata (subject, attribute, factual count, and CF count). We exclude binary yes/no questions, since VLMs show a systematic tendency to answer "yes" (Chan et al., 2026), which would confound our analysis. We therefore use two formats: (i) *four-option Multiple-Choice Questions (MCQ)*, where distractors are balanced around both the canonical and CF counts, and the option order is randomized to reduce positional bias (Atabuzzaman et al., 2025; Zeno et al., 2025); and (ii) *Open-Ended (OE) completion*, where the model completes a sentence with only the count and part name. Templates and full construction details are provided in Appendix B.1.

Human Study. We assess human performance on CounterCount to see how challenging the task is for humans, using human annotators recruited through Rapidata[2], a high-throughput platform for collecting human responses at scale. We use the MCQ format to minimize ambiguity and reduce

[1] https://deepmind.google/models/gemini-image/
[2] https://www.rapidata.ai/

free-form response errors. Each image is independently annotated by 25 qualified annotators, and since each image is displayed for a few seconds, this encourages quick perceptual judgments rather than deliberate analysis. The final answer is determined by majority voting. Full details of the human study are provided in Appendix D.

2.2 Region-Level Annotations

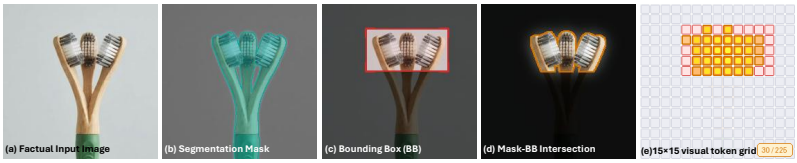


Figure 3: **Region-level annotation pipeline.**

In addition to paired factual-CF images, *CounterCount* provides region-level annotations for each edited instance, including segmentation masks and bounding boxes (BB). As shown in Figure 3, these annotations localize the edited object and the specific count-relevant attribute. Starting from the CF image (a), we use SAM (Carion et al., 2026) to obtain a segmentation mask for the full object containing the edit (b). We then manually annotate a BB around the modified attribute (c). The object mask captures the broader visual context, while the BB focuses on the localized structural change. Their intersection, denoted Mask-BB, isolates the pixels corresponding to the edited count-relevant region (d). We map the mask, BB, and Mask-BB regions onto the fixed visual-token grid of the vision encoder (e). This region-to-token alignment identifies both object-level tokens and localized anomaly tokens, enabling analyses and diagnosis targeted at the visual evidence required for counting rather than the image as a whole.

3 CounterCount Evaluation and Analysis

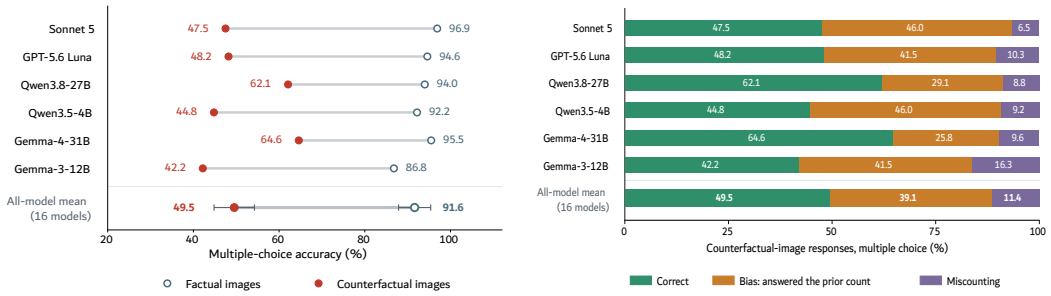
We evaluate VLMs on CounterCount's 554 image pairs using two question formats: MCQ and OE. We report MCQ results in the main paper, as this matches the format of our human study, and provide OE results in Appendix C.2.

Experimental Setup. We evaluate 16 VLMs: 14 open-weight models from six families, namely Qwen3-VL (4B, 8B, 32B) (Bai et al., 2025), Qwen3.5 (4B, 9B, 27B) (Qwen Team, 2026a), Qwen3.6 (27B) (Qwen Team, 2026b), Qwen3.8 (27B) (Qwen Team, 2026c), Gemma-3 (4B, 12B, 27B) (Kamath et al., 2025), and Gemma-4 (E2B, 12B, 31B) (Abd et al., 2026), and two closed-source models, GPT-5.6 Luna (OpenAI, 2026) and Sonnet 5 (Anthropic, 2026). Open-weight models run locally on a single NVIDIA A100 80GB GPU with greedy decoding, while closed-source models are accessed through their APIs. Unless otherwise stated, we disable reasoning and limit outputs to 128 tokens for all models. For clarity, figures in the main paper report six representative models: the two closed-source models, the two strongest open-weight models (Gemma-4-31B and Qwen3.8-27B), and a medium (Gemma-3-12B) and a small (Qwen3.5-4B) model to cover different scales, with results for all models in Appendix C. Analyses that require model internals or thinking mode use a subset of models, specified in the corresponding sections.

Evaluation Metrics. We extract numeric answers from model responses using regular expressions. When a response contains multiple numbers, we prompt the same VLM in its baseline configuration to identify the intended answer. On factual images, the canonical count is the correct answer, so we report accuracy only. On CF images, we classify each prediction as *correct* if it matches the edited count, *biased* if it matches the canonical count, and *other error* otherwise. We report ***accuracy*** as the fraction of correct predictions and ***bias*** as the fraction of biased predictions, which measures how often models favor the canonical count over the image content.

3.1 Baseline Results

Figure 4a presents MCQ results for the six selected models, along with the mean across all 16 models with 95% confidence intervals. Every model performs worse on CF images, with mean



(a) MCQ accuracy on factual and CF images.

(b) CF-image predictions split.

Figure 4: **Models' performance on CounterCount.**

accuracy dropping from 91.6% on factual images to 49.5% on CF images, a 42.1-point drop. In contrast, humans remain near ceiling on both (98.9% and 97.5%; Figure 1), leaving a 48-point gap on CF images.

Since the canonical count is also the correct answer on factual images, factual accuracy alone cannot reveal whether models count from the image or recall from prior knowledge. For example, Sonnet 5 achieves the highest factual accuracy (96.9%) but reaches only 47.5% on CF images, below the open-weight Gemma-4-31B (64.6%) and Qwen3.8-27B (62.1%).

Most CF errors are not random miscounts. As shown in Figure 4b, models return the canonical count in 39.1% of CF predictions, compared with only 11.4% for other errors, meaning that 77% of errors default to the canonical count. If failures stemmed from limited counting ability, errors would be spread across different counts rather than concentrated on the canonical one. This indicates that models mostly favor prior knowledge over visual evidence, rather than simply failing to count. OE results show the same trend, with mean accuracy falling from 90.9% on factual to 46.8% on CF images and a bias rate of 39.3%. Full MCQ and OE results are in Appendices C.1 and C.2.

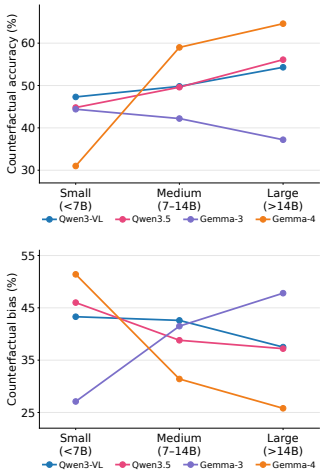


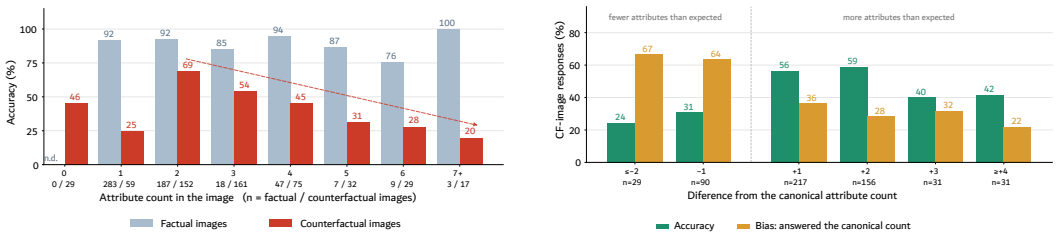
Figure 5: **Effect of model size.**

Does scaling improve performance? Within open-weight families, CF accuracy consistently increases and bias decreases with model size (Figure 5), with Gemma-3 as the exception, where accuracy declines and bias rises. However, even the best model remains far below its factual accuracy, and both closed-source models, despite their presumably larger scale, fall below the best open-weight models, suggesting that scale alone does not resolve the bias.

3.2 CounterCount Results Analysis

Are all categories equally difficult? Some categories are much harder than others (Appendix C.3). Bugs and mythical creatures have the lowest CF accuracy (20% and 29%), while housing reaches 82%, followed by flags and food (78%). Lower accuracy coincides with higher bias, as models return the canonical count for 73% of CF bug images, compared with only 9% for housing. Symbols stand out with the lowest factual accuracy (66%) and a smaller drop on CF images (54%), suggesting that models struggle with this category even before any modifications.

Does the attribute count matter? Consistent with prior work showing that VLM counting degrades as counts increase (Sengupta et al., 2025; Qharabagh et al., 2025; Alghisi et al., 2025), CF accuracy for counts of two or more falls from 69% at two to 20% at seven or more (Figure 6a), confirming our choice to keep most counts small. Factual accuracy, in contrast, varies across counts (76–100%) without a declining trend, since the canonical count yields the correct answer regardless of how



(a) (b)

Figure 6: **Effect of attribute count.** (a) MCQ accuracy on factual and CF images by attribute count. (b) CF accuracy and bias rate by the difference from the canonical count, with negative values indicating removal and positive values indicating addition.

many attributes there are. Counts of zero and one are exceptions, with low accuracy despite their small size. These cases mostly involve attribute removal, where models are far more likely to default to the canonical count. The bias rate reaches 64–67% for removal (*e.g.* 4 legs → 3 legs), compared with 22–36% for addition (*e.g.* 1 tail → 3 tails; Figure 6b). This suggests that missing attributes are easier to overlook than added ones, leading models to fill the gap with prior knowledge.

4 Understanding Model Behavior on Counterfactuals

Models show a surprising gap in counting capabilities between factual and CF images. In this section, we study this behavior more closely, showing that superimposing a System-1/2 (Kahneman, 2011) model partially correlates with model bias (Section 4.1), as well as the partial presence of necessary information in the activation space (Section 4.2).

4.1 Models Exhibit System 1-like Answer Behavior

We hypothesize that models exhibit System-1-like thinking (Kahneman, 2011) when answering questions, where they take a glance at the input image, recognize the object, and ultimately fall back on knowledge priors when answering questions about attributes. We study this hypothesis from four angles below.

4.1.1 Knowledge Priors are Caused by Visual Processing

A natural question is whether the explicit naming of the subject prompts models to return its canonical count purely based on knowledge. We ablate this by replacing subject names with generic phrasing (e.g., "parrot" → "animal"), rerunning our experimental validation (see Appendix C.4 for details), and find almost no change in accuracy or bias. This suggests that the source of the knowledge is not originating from the language alone, but must instead originate from the vision side, which we investigate in the following section.

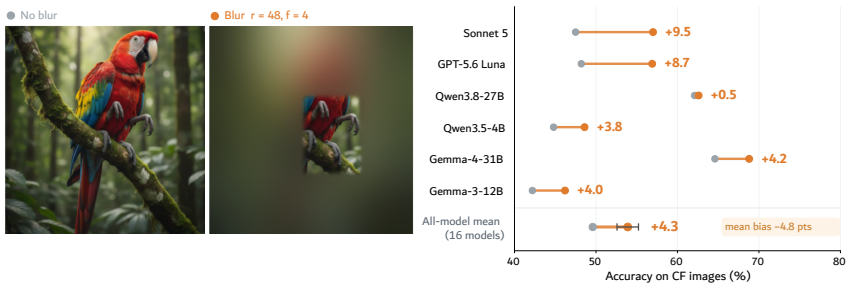


Figure 7: **Effect of visual context degradation.** Example image before and after blurring (left), alongside changes in CF-image MCQ accuracy (right).

4.1.2 DEGRADING VISUAL CONTEXT BY BLURRING

Models may recognize an object's overall shape and use its canonical count as a shortcut. Inspired by (Yang et al., 2023), we test whether degrading the visual context reduces this reliance. Using CounterCount's BB annotations, we keep the counted attribute regions sharp and blur everything else, including the rest of the subject and the background, so that the attributes remain countable while the subject becomes harder to recognize. We use a blur radius of 48 pixels and feathering of 4 pixels, selected through a parameter sweep on Qwen3-VL-8B (Appendix C.5). As shown in Figure 7, blurring improves mean CF MCQ accuracy by 4.3 points and reduces bias by 4.8 points across 16 models, with similar results in OE (Appendix C.5). These gains suggest that recognizing the subject is at least partially responsible for the bias. However, blurring cannot fully prevent recognition, since the attributes themselves often reveal the subject (*e.g.* the shape of a leg or feather suggests a bird). Removing more visual information would make the image ambiguous, for example, making it unclear whether the attributes belong to one subject or several.

4.1.3 CUEING THE MODEL THROUGH INSTRUCTIONS

Human observers typically benefit from *cueing* (Kahneman, 2011), where they receive a hint that the observed stimuli might be special. We therefore wonder if this applies to models as well, causing them to more carefully examine the visual information. We provide them with this instruction: *"Some images are digitally altered, so counts may be unrealistic. Count only what you actually see."*.

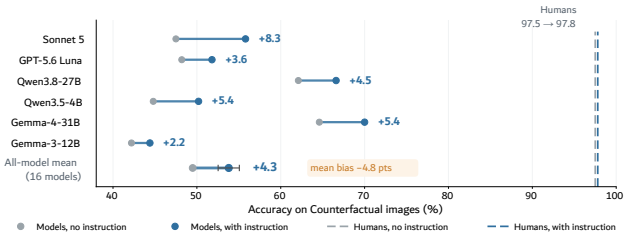


Figure 8: **Effect of cueing.** MCQ accuracy on CF images with and without the instruction, compared with humans.

On factual images, the instruction has little effect for either models or humans, as expected (Section C.6.1). On CF images, however, the two diverge. Models benefit from the cue, with mean MCQ accuracy increasing by 4.3 points and bias decreasing by 4.8 points, suggesting that giving instructions helps them rely more on visual evidence. Yet their mean accuracy reaches only 53.8%, still far below humans, whose accuracy remains nearly unchanged (97.5% without and 97.8% with the instruction; Figure 8). Since the task is simple and unambiguous, humans already count from the image and do not need the cue, whereas models improve only partially even when explicitly told that counts may be unusual. For the human evaluation, we follow the protocol of Section 2.1 with a new annotator pool validated with the instruction included, for consistency. Full OE and remaining MCQ results are provided in Appendix C.6.

4.1.4 THINKING MODES

What if models reason before answering? This most closely resembles System-2 processing (Kahneman, 2011). We enable thinking mode with each model's default sampling parameters and a maximum budget of 16,384 tokens for two model families that support it, Qwen3.5 (4B, 27B) and Gemma-4 (E2B, 31B), selecting a small and a large model from each to examine how the effect varies with size. As shown in Figure 9, thinking improves mean CF accuracy by 3.8 points in MCQ and 4.6 in OE (Figure 24), while reducing bias by 3.2 and 4.1 points, respectively. See Appendix C.7 for details.

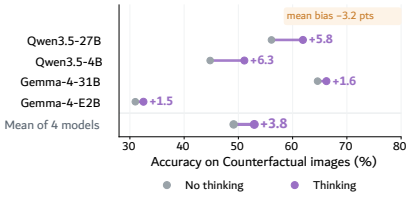


Figure 9: **Results with thinking.**

Takeaway. Models fall back to canonical counts upon a shallow recognition of the object, not semantic cues in the prompt. Encouraging System-2-like thinking by breaking visual context, cueing, or thinking mode reduces but does not fully alleviate bias.

4.2 MODELS HAVE THE INFORMATION INTERNALLY BUT UNDER-USE IT

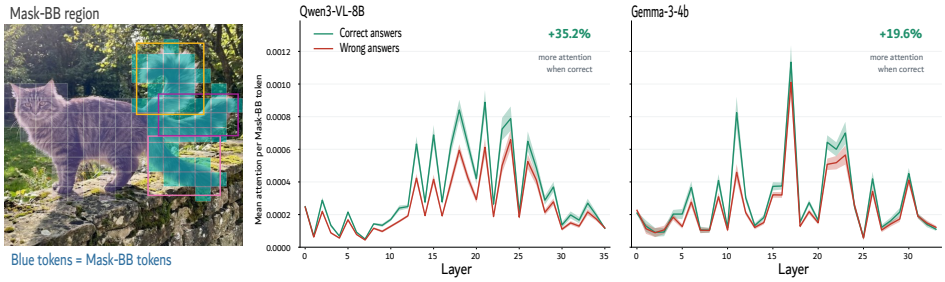


Figure 10: **Attention to count-relevant regions.** Mean attention per Mask-BB token across layers for correct and incorrect MCQ answers on CF images.

Beyond changing the input or the prompt, we ask whether models already have access to the visual evidence but fail to use it. Do models give enough attention to the relevant visual information? For each CF image, we measure the attention from the last token, which predicts the answer, to the Mask-BB tokens at each layer, averaged over heads and normalized by the number of selected tokens, and compare correct and incorrect answers. As shown in Figure 10, averaged across layers, correct answers receive 35.2% more attention in Qwen3-VL-8B and 19.6% more in Gemma-3-4B. Similar patterns appear in OE (Appendix C.8). This suggests that models may give insufficient attention to count-relevant regions when they answer incorrectly. We therefore test whether boosting attention to these regions can help recover correct answers.

Attention Modulation. Following prior work on hallucination in VLMs that modifies attention to visual tokens (Liu et al., 2024a; Zhao et al., 2025b; Jiang et al., 2025; Yu et al., 2025; Tu et al., 2025), we use a simple, unified formulation that amplifies, suppresses, or masks attention on selected heads through additive adjustments to attention logits. Let $z_{h,i,j}$ denote the attention logit from query position i to key position j in head h. For target visual tokens T and the remaining visual tokens B outside the target region, we apply

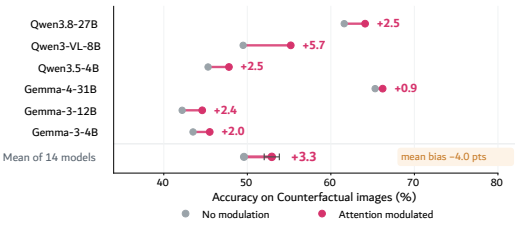


Figure 11: **Attention modulation results.**

$$\tilde{z}_{h,i,j} = \begin{cases} z_{h,i,j} + \log(\alpha), & j \in T \\ z_{h,i,j} + \log(\beta), & j \in B \end{cases} \tag{1}$$

where $\alpha \geq 1$ amplifies target attention and $0 \leq \beta \leq 1$ controls background suppression. A value of one leaves the logit unchanged, while zero masks the token by setting its logit to $-\infty$. We use Mask-BB tokens as the target and test two modes: target amplification alone ($\beta = 1$) and target amplification with background masking ($\beta = 0$). We rank attention heads using AUROC, computed as the fraction of token pairs in which a visual token inside the BB receives more attention than one outside it. Higher scores indicate stronger attention to the annotated region. We test 12 configurations combining these modes, $\alpha \in \{2, 3\}$, and the top 5%, 7%, or 10% of heads, selecting the best-performing configuration for each model as an upper-bound diagnostic. Full methodological and experimental details are provided in Appendix C.8.

As shown in Figure 11, attention modulation improves mean CF MCQ accuracy by 3.3 points and reduces bias by 4.0 points across 14 open-weight models, with similar gains in OE (Section C.8.5). Since modulation only reweights visual tokens the model already receives, this indicates that the evidence needed to count correctly is partially available but underused.

Takeaway. Models under-attend to count-relevant regions when they fail. Redirecting attention toward these regions, without adding any information, partially recovers correct answers, indicating that the visual evidence is available but underused in favor of prior knowledge.

5 Related Work

Language Priors and Bias in VLMs. Large Language Models (LLMs) exhibit systematic biases across demographic, cultural, and semantic dimensions (Sheng et al., 2019; Vo et al., 2025c; Parrish et al., 2022; Wang et al., 2024; Naous et al., 2024; Shin et al., 2024). Because Vision-Language Models (VLMs) rely on pretrained LLM backbones, these priors transfer to multimodal reasoning as spurious or stereotypical correlations (Hall et al., 2023; Raj et al., 2024; Howard et al., 2024; Yang et al., 2025; Ullman, 2024). Consequently, language priors can dominate visual evidence, leading models to produce prior-consistent answers even when the image supports a different response (Yang et al., 2025; Ullman, 2024; Deng et al., 2025; Lee et al., 2025; Vo et al., 2025a; Fu et al., 2025; Golovanevsky et al., 2025; Liu et al., 2025). Recent analyses suggest relevant visual information is often encoded in cross-modal representations but attenuated during generation by the language backbone (Tong et al., 2024; Zhao et al., 2025a; Vo et al., 2025b; Liu et al., 2024a; Favero et al., 2024; Li et al., 2026). This creates a modality imbalance: visual evidence is insufficiently used when it conflicts with linguistic expectations. We study this phenomenon through counting, placing canonical object knowledge (e.g., typical numbers of ears or wheels) in direct conflict with localized visual evidence.

Counterfactual Images for Diagnosing Bias. Counterfactual images test whether VLM predictions are grounded in vision or memorized priors (Lee et al., 2025; Wu et al., 2025; Ortu et al., 2025; Luo et al., 2025). Prior benchmarks often trigger hallucinations via prompt bias or semantically inconsistent scenes, which conflates multiple factors and complicates error attribution. Counting offers a sharper diagnostic because answers are objective and depend purely on visual enumeration (Paiss et al., 2023; Qharabagh et al., 2025; Rahmanzadehgervi et al., 2024; Sengupta et al., 2025; Campbell et al., 2024). However, VLMs can miscount even when attending to relevant objects (Alghisi et al., 2025; Hasani et al., 2024; Vo et al., 2025b; Rudman et al., 2025). Mechanistic counting studies exist (Che et al., 2026), but focus on standard images without prior conflicts. Closest to our setting, (Vo et al., 2025b) uses counterfactual images of familiar subjects to test canonical expectations. Yet, their limited edits (e.g., adding one leg), inclusion of high-count cases, and reliance on geometric shapes make it difficult to isolate prior-driven failures from standard perceptual limits (Rudman et al., 2025). A similar benchmark, CDH-Bench (Chen et al., 2026), studies the same prior-over-evidence failure but does not explicitly focus on counting, treating it as only one of three anomaly types, contains small counting pairs, includes ambiguous images, and lacks localized region annotations. We compare CounterCount with existing datasets in Table 2. CounterCount provides more image pairs, broader subject and attribute coverage, and is the only dataset with localized region annotations.

6 Conclusion

We introduced *CounterCount*, a diagnostic dataset for studying whether VLMs count from visual evidence or canonical priors. Across 16 VLMs, accuracy drops from 91.6% on factual images to 49.5% on CF images, while humans remain near the ceiling, and most errors return the canonical count. Our analysis reveals two underlying behaviors: (1) Models exhibit System-1-like behavior: correct answers on factual images say little about whether they looked closely, since knowing the subject is enough, so paired counterfactuals are needed to measure visual grounding. Breaking the visual context, cueing models that counts may be unusual, or letting them reason all lead them to rely more on the visual cues. (2) Models have the visual evidence but underuse it, as prior knowledge outweighs it. Redirecting their attention toward count-relevant regions shifts predictions toward the image. We release CounterCount to support the community in analyzing and understanding such failures, toward VLMs that ground their answers in visual evidence rather than prior knowledge.

Limitations. CounterCount uses synthetic images, as the controlled construction of factual-counterfactual pairs requires changing a subject's attribute count while holding the remaining visual content fixed–a condition that is exceedingly difficult to satisfy with naturally occurring images. We attempted to cover a broad range of subjects and attributes across 17 categories, yet other subject types with strong canonical priors remain to be explored. Finally, the hyperparameter search for our attention modulation was limited, and other configurations may have had even stronger debiasing effects.

AI USE STATEMENT

In this work, we used generative AI tools to generate all images in *CounterCount*. Factual and counterfactual images were produced with Nano-Banana, and subject masks were obtained with SAM. Every image and mask was manually verified by the authors for photorealistic quality, semantic coherence, correctness of the target count, and accurate masking of the subjects. Bounding boxes of the counted attributes and all image metadata were annotated manually by the authors. We have not used generative AI tools to generate research ideas, design experiments, or write questions. MCQ and OE questions were created from templates using the manually written metadata. Additionally, we used generative AI tools as coding assistants to make minor modifications to experiment scripts (*e.g.* enabling thinking mode), monitor experiment runs, and write plotting code for figures. We also used them to assist with literature search and to polish the writing of the manuscript. We have reviewed all AI-assisted work. All AI-modified code was checked by the authors and its outputs verified against our raw results. Papers surfaced by AI-assisted search were read by the authors, and only those relevant to this work were cited. All AI-polished text was checked for accuracy against our experiments and results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We release CounterCount, including all factual and counterfactual images, metadata, subject masks, and bounding-box annotations, through an anonymous link provided in the supplementary materials (see Appendix A.3). The dataset construction and annotation process is described in Section 2, and the human study protocol, including annotator validation, is described in Section 2.1. The evaluated models, decoding settings, and evaluation metrics are specified in Section 3, with question templates and answer extraction details in Appendix B.1. Implementation details for each analysis are provided in the appendix, including the blur parameter sweep (Appendix C.5), cueing instruction and human evaluation (Appendix C.6), thinking-mode settings (Appendix C.7), and attention analysis and modulation configurations (Appendix C.8). All open-weight models are evaluated with greedy decoding unless otherwise stated, and closed-source models are accessed through their APIs.

REFERENCES

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-

Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Simone Alghisi, Gabriel Roccabruna, Massimo Rizzoli, Seyed Mahed Mousavi, and Giuseppe Riccardi. [de—re]constructing vlms' reasoning in counting, 2025. URL https://arxiv.org/abs/2510.19555.

Duy Le Dinh Anh, Patrick Amadeus Irawan, and Tuan Van Vo. Counting to four is still a chore for vlms, 2026. URL https://arxiv.org/abs/2604.10039.

Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, June 2026. Accessed: 2026-09-13.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. Benchmarking and mitigating mcqa selection bias of large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.1703.pdf.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. Understanding the limits of vision language models through the lens of the binding problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://neurips.cc/virtual/2024/poster/95266.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Tsan Tsai Chan, Varsha Suresh, Anisha Saha, Michael Hahn, and Vera Demberg. System-mediated attention imbalances make vision-language models say yes, 2026. URL https://arxiv.org/abs/2601.12430.

Liwei Che, Zhiyu Xue, Yihao Quan, Benlin Liu, Zeru Shi, Michelle Hurst, Jacob Feldman, Ruixiang Tang, Ranjay Krishna, and Vladimir Pavlovic. Counting circuits: Mechanistic interpretability of visual reasoning in large vision-language models, 2026. URL https://arxiv.org/abs/2603.18523.

Kesheng Chen, Yamin Hu, Qi Zhou, Zhenqian Zhu, and Wenjian Luo. Cdh-bench: A commonsense-driven hallucination benchmark for evaluating visual fidelity in vision-language models, 2026. URL https://arxiv.org/abs/2603.27982.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, June 2025.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, June 2024.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations, 2025. URL https://arxiv.org/abs/2506.08008.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524. URL https://aclanthology.org/2024.cl-3.8/.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://neurips.cc/virtual/2023/poster/73666.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Understanding counting mechanisms in large language and vision-language models. *arXiv preprint arXiv:2511.17699*, 2024.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11975–11985, June 2024.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, June 2025.

Daniel Kahneman. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, NY, US, 2011.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael

Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-bench: Measuring language priors in large vision-language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.231. URL https://aclanthology.org/2025.findings-naacl.231/.

Jiaang Li, Chengzu Li, Zhaochong An, Yifei Yuan, Xi Liu, Serge Belongie, and Vésteinn Snæbjarnarson. Seeing or knowing? visual context sensitivity in multimodal large language models. *arXiv preprint arXiv:2607.26326*, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms, 2024a. URL https://eccv.ecva.net/virtual/2024/poster/2599.

Yuan 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? In *The European Conference on Computer Vision (ECCV)*, 2024b.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms, 2025. URL https://arxiv.org/abs/2510.17771.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://iclr.cc/virtual/2024/oral/19768.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix

Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/luo25b.html.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16366–16393, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.862. URL https://aclanthology.org/2024.acl-long.862/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed: 2026-09-13.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=GhSp6uvUuR.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3170–3180, October 2023.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Muhammad Fetrat Qharabagh, Mohammadreza Ghofrani, and Kimon Fountoulakis. LVLM-COUNT: Enhancing the counting ability of large vision-language models, 2025. URL https://openreview.net/forum?id=GsCMKwyfWm.

Qwen Team. Qwen3.5: Towards native multimodal intelligence. https://qwen.ai/blog?id=qwen3.5, February 2026a. Accessed: 2026-09-10.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. https://qwen.ai/blog?id=qwen3.6-27b, April 2026b. Accessed: 2026-09-10.

Qwen Team. Qwen3.8: Large language model series. https://github.com/QwenLM/Qwen3.8, 2026c. Accessed: 2026-09-10.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, December 2024.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. BiasDora: Exploring hidden biased associations in vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10439–10455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.611. URL https://aclanthology.org/2024.findings-emnlp.611/.

William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. Forgotten polygons: Multimodal large language models are shape-blind. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11983–11998, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.620. URL https://aclanthology.org/2025.findings-acl.620/.

Saurav Sengupta, Nazanin Moradinasab, Jiebei Liu, and Donald E. Brown. Can vision-language models count? a synthetic benchmark and analysis of attention-based interventions. *arXiv preprint arXiv:2511.17722*, 2025.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1339. URL https://aclanthology.org/D19-1339/.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16122–16143, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.954. URL https://aclanthology.org/2024.findings-acl.954/.

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)*, pp. 9568–9578, June 2024.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *Int. J. Comput. Vision*, 134(1), December 2025. ISSN 0920-5691. doi: 10.1007/s11263-025-02607-z. URL https://doi.org/10.1007/s11263-025-02607-z.

Tomer Ullman. The illusion-illusion: Vision language models see illusions where there are none, 2024. URL https://arxiv.org/abs/2412.18613.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased: Counting legs of an animal is surprisingly hard. In *2nd AI for Math Workshop @ ICML 2025*, 2025a. URL https://icml.cc/virtual/2025/52429.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased, 2025b. URL https://openreview.net/forum?id=4GWfYyo6FS.

An Vo, Mohammad Reza Taesiri, Daeyoung Kim, and Anh Totti Nguyen. B-score: Detecting biases in large language models using response history. In *Forty-second International Conference on Machine Learning*, 2025c. URL https://icml.cc/virtual/2025/poster/44236.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6349–6384, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

doi: 10.18653/v1/2024.acl-long.345. URL https://aclanthology.org/2024.acl-long.345/.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, Junjie Ao, Enyan Dai, Lei Feng, Xiang Zhang, and Suhang Wang. Lanp: Rethinking the impact of language priors in large vision-language models, 2025. URL https://arxiv.org/abs/2502.12359.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24993–25006. Curran Associates, Inc., 2023. doi: 10.52202/075280-1086. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4e9fa6e716940a7cfc60c46e6f702f52-Paper-Conference.pdf.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Zhe Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://neurips.cc/virtual/2025/loc/san-diego/poster/121529.

Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-grounded dual-path attention intervention for object hallucination mitigation in lvlms, 2025. URL https://arxiv.org/abs/2511.09018.

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. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 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)*, pp. 9556–9567, June 2024.

Giselle Zeno, Nour Jedidi, and Steven R. Gomez. Choosing 'right' from wrong: A closer look at selection bias in spatial multiple-choice questions in large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 535–544, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/BEAM/html/Zeno_Choosing_Right_from_Wrong_A_Closer_Look_at_Selection_Bias_CVPRW_2025_paper.html.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5625–5644, August 2024. ISSN 0162-8828. doi: 10.1109/TPAMI.2024.3369699. URL https://doi.org/10.1109/TPAMI.2024.3369699.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19666–19690, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.995. URL https://aclanthology.org/2025.emnlp-main.995/.

Jianfei Zhao, Feng Zhang, Xin Sun, Chong Feng, and Zhixing Tan. Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention, 2025b. URL https://arxiv.org/abs/2511.20032.

Overview of the Appendix

A Additional Details on CounterCount Data **19**

A.1 Dataset Statistics 19

A.2 Dataset Comparison 19

A.3 Factual–Counterfactual pairs 20

B Details on Prompt Design **20**

B.1 Distractor Generation for MCQs 20

B.2 VLM-as-a-Judge for Answer Extraction 21

B.3 Open-Ended and MCQ Prompts 21

C Additional Experimental Results **21**

C.1 Multiple-Choice Baseline Results 21

C.2 Open-ended Baseline Results 22

C.3 Categories Results 23

C.4 Generic vs. Specific subjects 23

C.5 Blurring Results 24

C.6 Instruction Results 25

C.7 Thinking 26

C.8 Attention 27

D Human Study Details **29**

D.1 Platform and Audience Construction 29

D.2 Validation Set 30

D.3 Annotation Protocol 30

D.4 Instruction Experiment: Audience Construction 30

A Additional Details on CounterCount Data

A.1 Dataset Statistics

Table 1: Statistics of the CounterCount dataset. The sample CF shows a representative CF instance (factual count → CF count). Total subjects and attribute types are counted as unique values across the whole dataset, so they are lower than the sum of the per-category columns.

Category	Pairs	Subjects	Attr. Types	Sample CF
Mammals	39	32	12	giraffe (1 head → 5)
Birds	31	27	6	parrot (2 legs → 4)
Crawling Creatures	31	19	15	spider (8 legs → 4)
Reptiles	30	21	12	gecko (4 legs → 1)
Sea Creatures	30	20	15	narwhal (1 tusk → 3)
Ancient Creatures	40	15	18	mammoth (1 trunk → 2)
Human	35	7	21	hand (5 fingers → 7)
Mythical Creatures	31	17	11	unicorn (1 horn → 4)
Fictional Characters	32	18	26	hello kitty (2 eyes → 1)
Functional Obj.	30	29	25	T-shirt (2 sleeves → 4)
Housing	37	33	27	cottage (1 chimney → 3)
Transportation	32	27	26	car (4 wheels → 6)
Landmarks	30	23	24	Fuji mountain (1 peak → 4)
Food	30	30	21	pineapple (1 crown → 3)
Flags	35	18	21	Japan flag (1 red circle → 2)
Logos	31	20	25	Starbucks logo (1 crown → 2)
Currency Sym.	30	23	9	naira (2 lines → 3)
Total	**554**	**379**	**228**	

A.2 Dataset Comparison

Table 2: Comparison of CounterCount with prior counterfactual, hallucination, and counting-oriented benchmarks. All CounterCount images are human-verified. *Counts estimated from reported image totals; exact numbers not confirmed in the source papers. **Unique images only; the original benchmark's totals also include multiple resolutions of each image.

Benchmark	Imgs / Pairs	Local. Annot.	Gen. Method	Question Type	Subj. / Attr.	Cat.	Conflict Type	Count. Task
VLind-Bench (Lee et al., 2025)	2576 / ∼1288*	✗	Auto	Binary	—	—	Language prior	✗
ViLP (Luo et al., 2025)	900 / —	✗	Auto	Open	—	—	Visual language prior	✗
VLMs are Biased (Vo et al., 2025a)	∼464** / —	✗	Semi-auto	Open/Binary	— / ∼19*	7	World knowledge	✓
HallusionBench (Guan et al., 2024)	181 / 165	✗	Manual	Binary	—	—	Lang. prior + illusion	✗
MMVP (Tong et al., 2024)	300 / 150	✗	Manual	MCQ/Binary	—	—	CLIP-blind perception	✗
LanP (Wu et al., 2025)	170 / —	✗	Manual	Binary	—	4	Visual ambiguity	Partial
CDH-Bench (Chen et al., 2026)	600 / 300	✗	Manual	MCQ/Binary	66 / 60	3	Object-level knowledge	Partial
CounterCount (ours)	**1003 / 554**	**Mask+BB**	**Manual**	**MCQ/Open**	**379 / 228**	**17**	**Object-level knowledge**	✓

A.3 FACTUAL–COUNTERFACTUAL PAIRS

Figure 12 shows representative factual and counterfactual examples drawn from across the seventeen categories in CounterCount[3]. Each pair preserves viewpoint, lighting, background, and object identity, and differs only in the count of a single structural attribute, illustrating the range of canonical priors the dataset places in conflict with visual evidence.

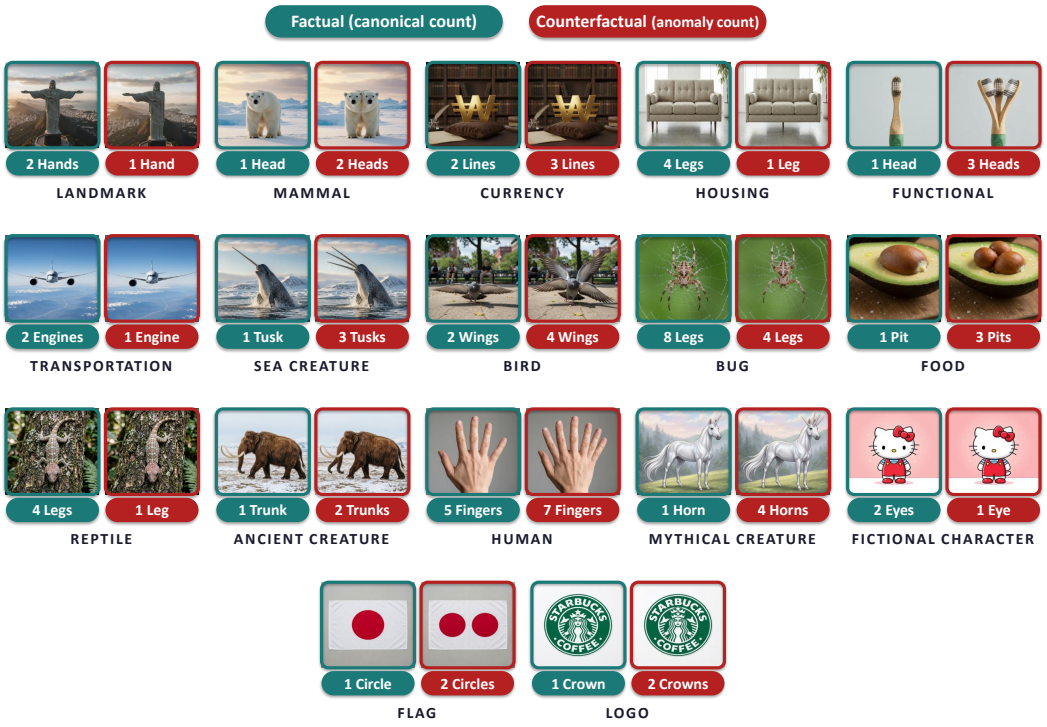


Figure 12: **Representative factual–counterfactual pairs from CounterCount, spanning all 17 semantic categories.** Each column pair shows one instance: **left image** (teal border) shows the corresponding factual image with canonical counts intact, while **right image** (red border) is the counterfactual in which a canonical attribute count is deliberately violated.

B DETAILS ON PROMPT DESIGN

B.1 DISTRACTOR GENERATION FOR MCQS

Table 3: Examples of distractor generation for MCQ options. c_o and c_a denote the *ordinary* and *anomalous* counts, respectively.

c_o	c_a	Gap	d_1	d_2	Generated Options
2	4	2	midpoint (3)	outside (5)	{2, 3, 4, 5}
2	7	5	midpoint (4)	outside (8)	{2, 4, 7, 8}
7	2	5	midpoint (4)	outside (8)	{2, 4, 7, 8}
2	3	1	below min (1)	outside (4)	{1, 2, 3, 4}
1	2	1	outside (3)	fallback (4)	{1, 2, 3, 4}
3	0	3	midpoint (1)	outside (4)	{0, 1, 3, 4}
1	0	1	outside (2)	fallback (3)	{0, 1, 2, 3}

Each multiple-choice question consists of four options: the correct anomalous count, the ordinary count, and two distractors. Distractors are generated programmatically based on the relationship between the ordinary count c_o and the anomalous count c_a.

[3] https://kaggle.com/datasets/c4cccfd90ef699f8bf315b71e23d08466bfe4fadcb59d77b65e263e41465ae45

When the gap between the two counts is greater than one, the first distractor is set to the midpoint $\lfloor (c_o + c_a)/2 \rfloor$, providing a plausible in-between value. The second distractor is set to one above the maximum of the two counts, serving as an out-of-range option. When the gap equals one, no meaningful midpoint exists, so the first distractor is instead set to one below the minimum of the two counts. If this yields an invalid value (i.e., zero or below), a fallback assigns the distractor to two above the maximum. This fallback also applies when $c_a = 0$, where the minimum is already zero and no smaller valid count exists. In all cases, duplicate options are resolved by incrementing the conflicting value until a unique candidate is found. Finally, all options are converted to word form (e.g., *three*) and shuffled, to prevent position bias, with a fixed per-image random seed, to ensure reproducibility. Table 3 illustrates the distractor generation strategy across representative cases.

B.2 VLM-AS-A-JUDGE FOR ANSWER EXTRACTION

We employ a VLM-as-a-judge approach to extract the final numeric answer from each response when multiple numbers are present. Specifically, we prompt the judge model with the following:

> *Question: {question}*
> *Model response: '{response}'*
> *The correct answer to this question is {correct_answer}.*
> *What single number did the model give as its final answer to the question? Reply with only a single digit, nothing else.*

The judge is provided with the original question, the full model response, and the ground-truth answer as context, and is asked to identify the single digit the evaluated model committed to as its final answer. This allows us to determine whether the model response was correct, biased, or a miscount.

B.3 OPEN-ENDED AND MCQ PROMPTS

Each image is evaluated under two prompting formats: OE and MCQ. Both formats share the same base question, which is constructed from the instance metadata using the object name and anomalous part:

> *How many {anomaly}s does this {name} have?*

For example, for a parrot with an anomalous number of legs, the base question becomes: *How many legs does this parrot have?*

In the OE format, the model is prompted to complete a sentence with just the count and the name of the part:

> *How many {anomaly}s does this {name} have? Complete the following sentence with just the count and the name of the part: The {name} has*

In the MCQ format, four options are appended to the base question, and the model is instructed to select one:

> *How many {anomaly}s does this {name} have? Choose one of the following options: {options}. Reply with only one word from the given options and nothing else.*

The MCQ options consist of the correct anomalous count, the ordinary count, and two distractors generated as described in B.1.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 MULTIPLE-CHOICE BASELINE RESULTS

Figure 13 shows baseline MCQ results for the models not included in the main paper.

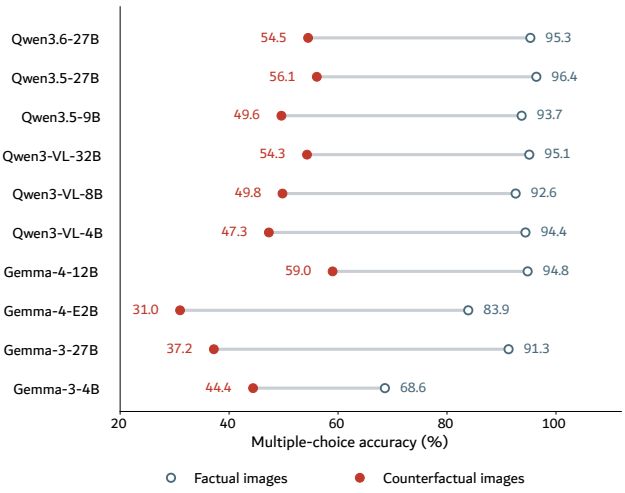


Figure 13: MCQ accuracy on factual and CF images.

C.2 Open-ended Baseline Results

Figure 14 shows baseline OE results for all evaluated models.

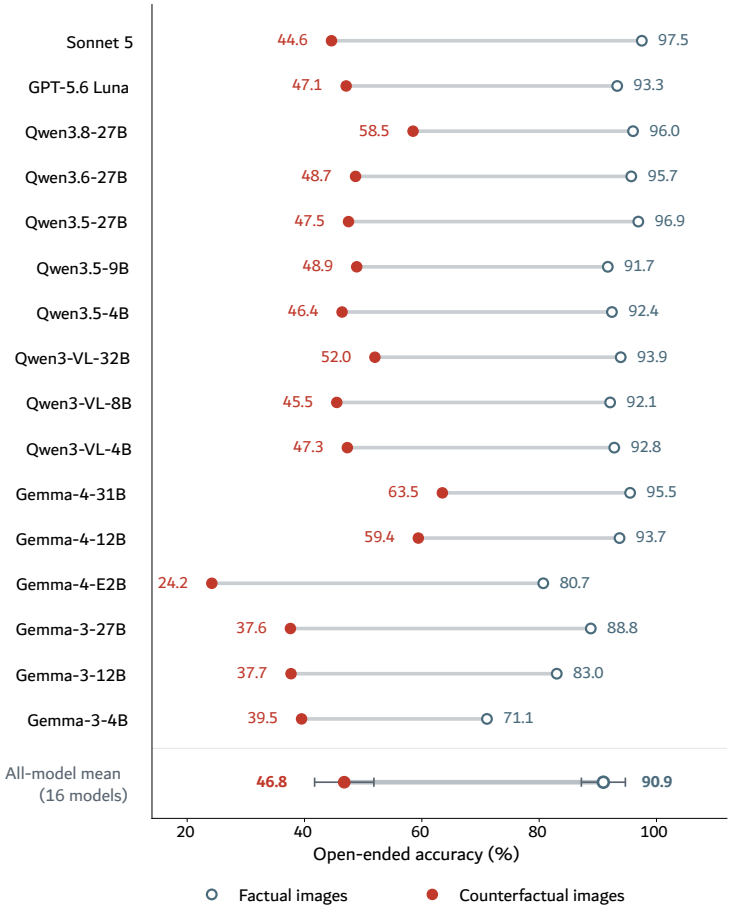


Figure 14: OE accuracy on factual and CF images.

C.3 Categories Results

Figure 15 reports factual and CF accuracy per category, and Figure 16 breaks down CF predictions into correct, biased, and other errors.

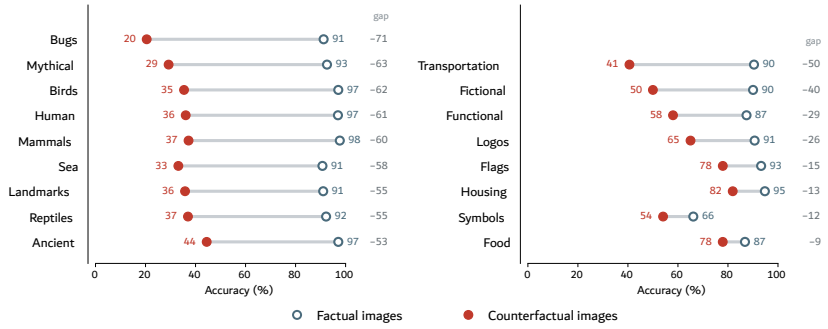


Figure 15: **Accuracy by category.** Factual and CF accuracy per category, averaged over MCQ and OE.

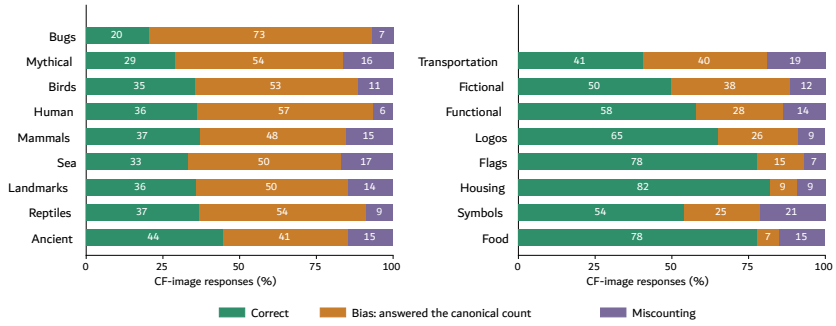


Figure 16: **Prediction breakdown by category.** Share of correct, biased, and other-error predictions on CF images per category, averaged over MCQ and OE.

C.4 Generic vs. Specific Subjects

To test whether naming the subject triggers the prior, we replace each specific subject name with a generic term assigned per category, leaving the rest of the question unchanged (*e.g.* "this parrot" $\rightarrow$ "this animal"). We use *animal* for Ancient Creatures, Birds, Bugs, Mammals, Reptiles, and Sea Creatures; *object* for Functional, Housing, and Transportation; *character* for Fictional Characters; *creature* for Mythical Creatures; and the singular category name for the remaining categories (*flag*, *food*, *human*, *landmark*, *logo*, and *symbol*). As shown in Figure 17, CF accuracy and bias remain unchanged in both formats.

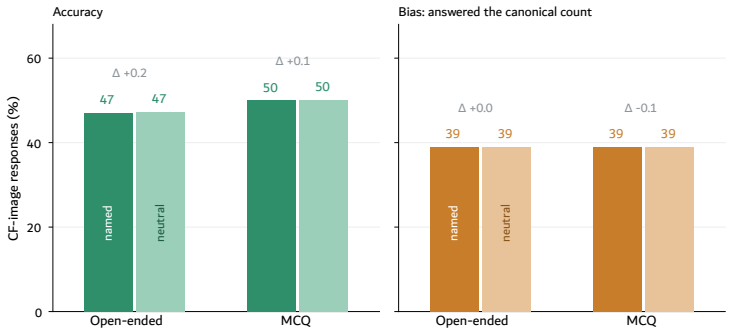


Figure 17: **Effect of subject naming.** CF accuracy (left) and bias rate (right) with specific and generic subject names, averaged over all models, for OE and MCQ.

C.5 Blurring Results

C.5.1 MCQ Results

Figure 18 shows MCQ results for the models not included in the main paper.

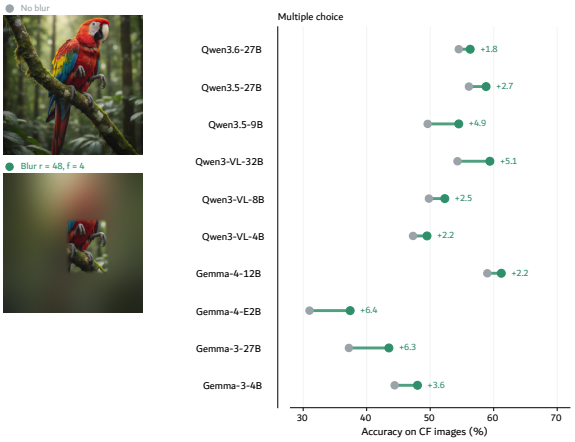


Figure 18: **Blurring (MCQ).** CF accuracy with and without blurring.

C.5.2 OE Results

Figure 19 shows OE results for all evaluated models.

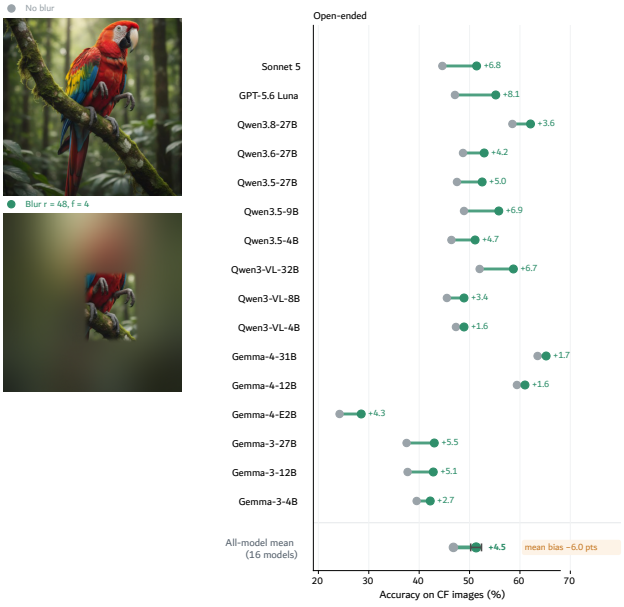


Figure 19: **Blurring (OE).** CF accuracy with and without blurring.

C.5.3 Blurring Sweep

Figure 20 shows CF accuracy and bias rate, averaged over MCQ and OE, across blur radii and feathering values on Qwen3-VL-8B. We select the setting with the highest accuracy (radius 48, feathering 4). Larger radii slightly reduce bias but do not improve accuracy, and since the sweep is performed on a single model, we avoid stronger blurring that may degrade performance on other models.

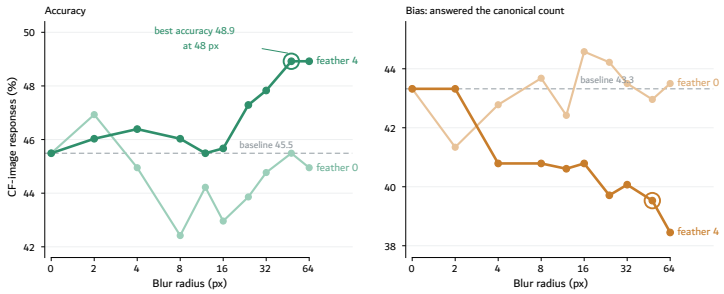


Figure 20: **Blur parameter sweep on Qwen3-VL-8B.** CF accuracy and bias rate, averaged over MCQ and OE, across blur radii and feathering values.

C.6 Instruction Results

C.6.1 Humans vs. Models

Figure 21 compares humans and models on factual and CF images with and without the instruction. On factual images, the instruction has little effect for either, while on CF images, models improve and humans remain near the ceiling.

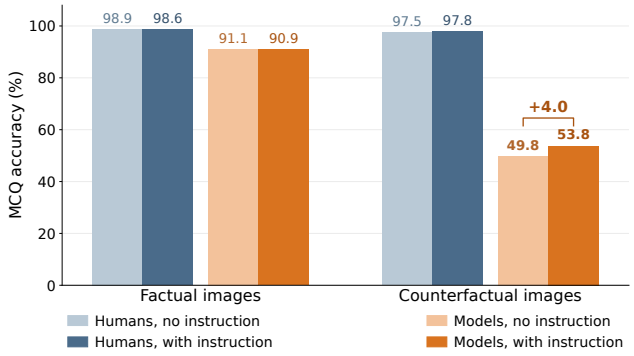


Figure 21: **Humans vs. models with cueing.** MCQ accuracy on factual and CF images with and without the instruction.

C.6.2 MCQ Results

Figure 22 shows MCQ results for models not included in the main paper.

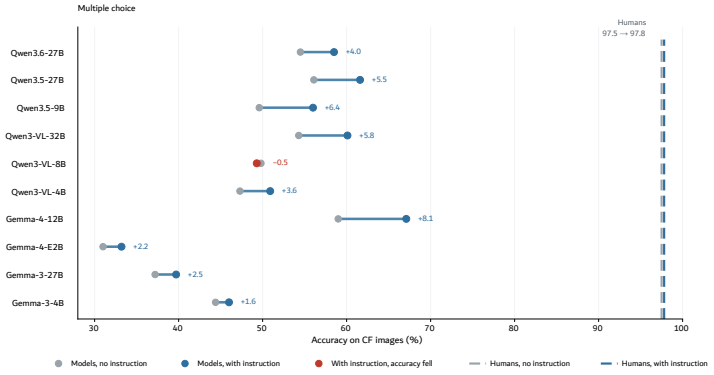


Figure 22: **Cueing (MCQ).** CF accuracy with and without the instruction for models not shown in Figure 8, compared with humans.

C.6.3 OE Results

Figure 23 shows OE results for all models.

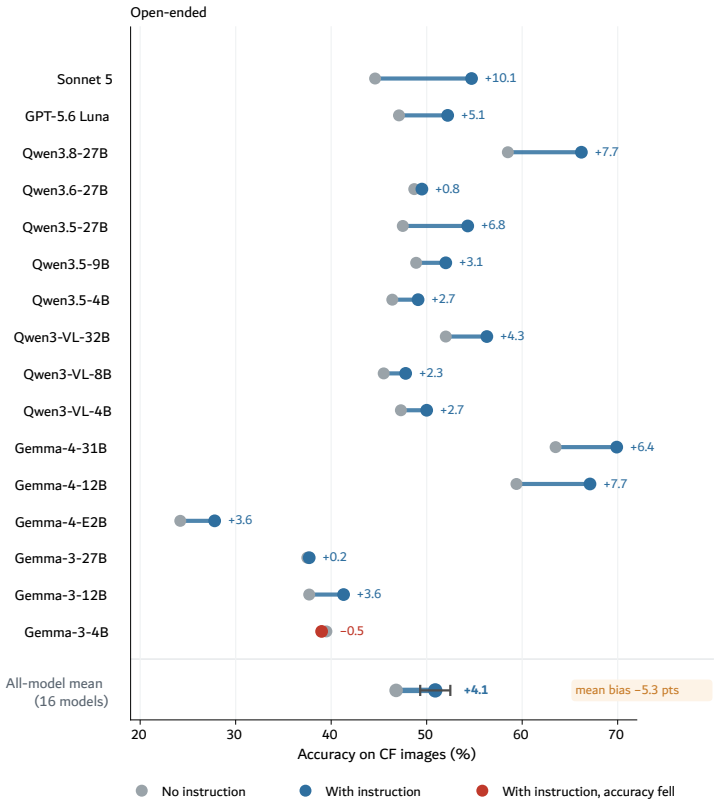


Figure 23: **Cueing (OE).** CF accuracy with and without the instruction for all models.

C.7 Thinking

C.7.1 OE Results

Figure 24 shows OE results with and without thinking for the four models.

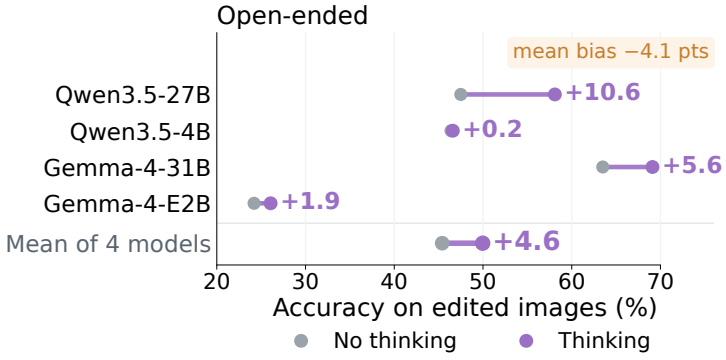


Figure 24: **Thinking (OE).** CF accuracy with and without thinking mode for Qwen3.5 (4B, 27B) and Gemma-4 (E2B, 31B).

C.7.2 Thinking Mode Decoding Details

We enable thinking mode through each model's chat template and keep the prompt unchanged from the baseline. Generation uses sampling with a maximum budget of 16,384 tokens and each model's default sampling parameters (Table 4). Qwen3.5-4B does not ship a generation configuration and therefore uses the library defaults, which are more diffuse than those of Qwen3.5-27B. After generation, we remove the reasoning block and extract the answer from the remaining response, following the procedure in Section 3. Since decoding is stochastic and results come from a single run, per-model numbers may vary across runs.

Table 4: **Sampling parameters for thinking mode.**

Model	Temperature	Top-k	Top-p
Qwen3.5-4B[†]	1.0	50	1.0
Qwen3.5-27B	0.6	20	0.95
Gemma-4-E2B	1.0	64	0.95
Gemma-4-31B	1.0	64	0.95

[†]No generation configuration provided; library defaults used.

C.8 Attention

C.8.1 Attention Analysis

For OE, where the answer spans multiple tokens, we follow the procedure in Section 4.2 and additionally average attention over the generated answer tokens. As shown in Figure 25, correct answers similarly receive more attention to count-relevant regions than incorrect ones.

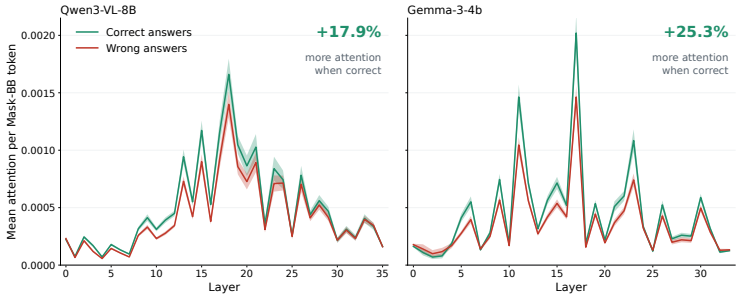


Figure 25: **Attention to count-relevant regions (OE).** Mean attention per Mask-BB token across layers for correct and incorrect OE answers on CF images.

C.8.2 Attention Modulation

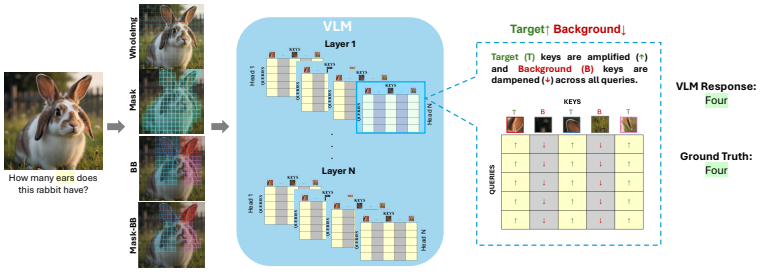


Figure 26: **Inference-time attention modulation.** The formulation supports modulating attention toward different visual-token regions (*WholeImg*, *Mask*, *BB*, and *Mask-BB*); we use Mask-BB throughout. We illustrate the T↑B↓ setting, which amplifies count-relevant target tokens while suppressing tokens outside the target.

The additive logit adjustment in Equation (1) corresponds to rescaling attention weights. Let $z_{h,i,j}$ denote the attention logit for head h from query position i to key position j, with standard attention

weights

$$A_{h,i,j} = \frac{\exp(z_{h,i,j})}{\sum_{k=1}^{n} \exp(z_{h,i,k})}. \tag{2}$$

Let $\Gamma \in \mathbb{R}^n$ with $\Gamma_k = \alpha$ if key k is targeted and $\Gamma_k = 1$ otherwise. Scaling the exponentiated logit at a selected position j yields

$$\tilde{A}_{h,i,j} = \frac{\alpha \exp(z_{h,i,j})}{\sum_{k=1}^{n} \Gamma_k \exp(z_{h,i,k})} = \frac{\exp(z_{h,i,j} + \log \alpha)}{\sum_{k=1}^{n} \exp(z_{h,i,k} + \log \Gamma_k)}. \tag{3}$$

Thus, amplification ($\alpha > 1$), suppression ($0 < \alpha < 1$), masking ($\alpha = 0$), and identity ($\alpha = 1$) are special cases of the same operation. When α is applied uniformly across all keys, it cancels in the softmax and leaves the attention distribution unchanged, so the modulation takes effect only when applied selectively. Given CounterCount's annotations, the target can be the whole image, the subject mask, the bounding boxes, or their overlap (Mask-BB); we use Mask-BB throughout.

C.8.3 Selected Configurations

Table 5 lists the configuration selected for each model from the 12 described in Section 4.2, along with the resulting gain in CF accuracy. All 14 models improve in both formats. Qwen models mostly favor amplification with outside masking, while Gemma models mostly favor amplification alone, suggesting that the optimal mode depends on the model family.

Table 5: **Selected attention modulation configuration per model.** Amp.: target amplification alone ($\beta = 1$); Amp.+Mask: target amplification with outside masking ($\beta = 0$). Gain: change in CF accuracy (points).

	MCQ				OE			
Model	Mode	α	Heads	Gain	Mode	α	Heads	Gain
Qwen3-VL-4B	Amp.	3	10%	+5.05	Amp.	3	7%	+2.35
Qwen3-VL-8B	Amp.+Mask	3	7%	+5.78	Amp.+Mask	3	5%	+5.42
Qwen3-VL-32B	Amp.+Mask	3	5%	+5.23	Amp.+Mask	2	7%	+5.60
Qwen3.5-4B	Amp.+Mask	2	7%	+2.53	Amp.+Mask	2	7%	+3.25
Qwen3.5-9B	Amp.+Mask	3	5%	+4.51	Amp.+Mask	2	5%	+2.89
Qwen3.5-27B	Amp.+Mask	3	10%	+2.89	Amp.+Mask	2	10%	+2.53
Qwen3.6-27B	Amp.	3	7%	+1.81	Amp.+Mask	2	5%	+1.26
Qwen3.8-27B	Amp.+Mask	3	7%	+2.53	Amp.+Mask	2	10%	+1.81
Gemma-3-4B	Amp.	3	10%	+1.99	Amp.	2	7%	+2.35
Gemma-3-12B	Amp.	2	5%	+2.35	Amp.	3	10%	+2.89
Gemma-3-27B	Amp.+Mask	2	10%	+3.79	Amp.+Mask	2	7%	+1.81
Gemma-4-E2B	Amp.	3	5%	+5.23	Amp.	3	5%	+2.53
Gemma-4-12B	Amp.	3	5%	+1.99	Amp.	2	5%	+2.53
Gemma-4-31B	Amp.	2	7%	+0.90	Amp.	2	10%	+0.72
Mean				**+3.33**				**+2.71**

C.8.4 MCQ Results

Figure 27 shows MCQ results for the eight open-weight models not included in the main paper.

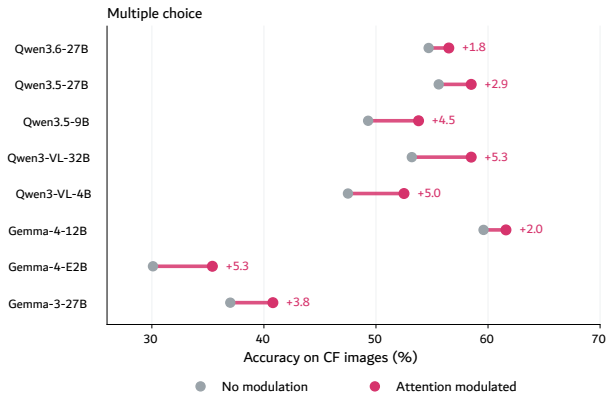


Figure 27: **Attention modulation (MCQ).** CF accuracy before and after modulation.

C.8.5 OE Results

Figure 28 shows OE results for all 14 open-weight models.

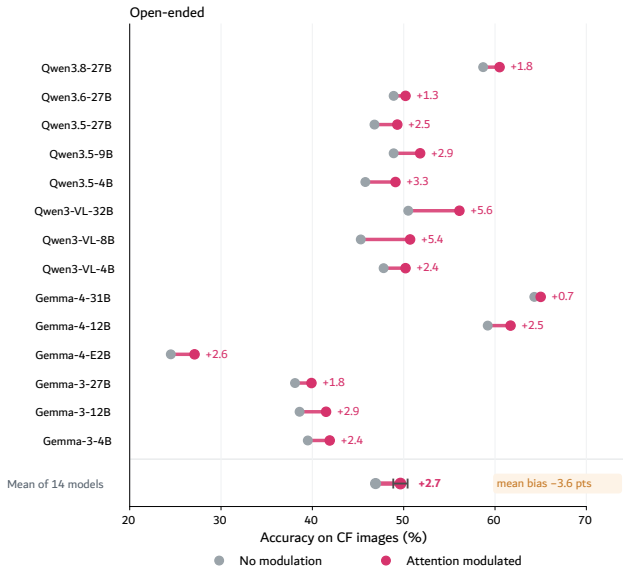


Figure 28: **Attention modulation (OE).** CF accuracy before and after modulation.

D Human Study Details

D.1 Platform and Audience Construction

All human annotations were collected through Rapidata[4], a high-throughput crowdsourcing platform. It displays each image for at most 25 seconds. Rather than drawing from Rapidata's open global pool, we constructed a *qualified custom audience* so that careless or inattentive annotators were screened out before the real counting task.

Qualification trial. Each candidate annotator was presented with a 10-image validation set consisting of 4 factual images (canonical counts) and 6 CF images (anomalous counts), all with known ground-truth answers. Questions were posed in the same four-option multiple-choice (MCQ) format used in the main evaluation. To be an annotator, a participant had to solve at least 8 of these 10 trial questions at $\geq$87.5% accuracy (`min_tasks`=8, `target_accuracy`=0.875).

[4] https://www.rapidata.ai/

Threshold justification. We deliberately set the annotator threshold at 0.875 rather than requiring a perfect score of 1.0. A perfect threshold would have wrongly rejected approximately one-fifth of otherwise excellent annotators who happened to make a single error over eight trials. The sub-perfect threshold therefore balances annotator quality against pool size, retaining strong counters while filtering out genuinely careless respondents.

D.2 Validation Set

Table 6 lists the 10 images used in the qualification trial. Each row shows the subject, the counted attribute, the count shown in the image, the canonical (expected) count, and whether the image is factual or CF. The corresponding images are shown in Figure 29.

Table 6: Qualification validation set. Images where the shown count equals the canonical count are factual; all others are CF. The set contains 4 factual and 6 CF images.

Subject	Attribute	MCQ Question	Shown	Canonical	Type
cat	tail	How many tails does this cat have?	1	1	Factual
fly	leg	How many legs does this fly have?	6	6	Factual
dog	ear	How many ears does this dog have?	2	2	Factual
whale	side flipper	How many side flippers does this whale have?	2	2	Factual
cat	leg	How many legs does this cat have?	3	4	CF
pikachu	red cheek	How many red cheeks does this pikachu have?	0	2	CF
pigeon	wing	How many wings does this pigeon have?	3	2	CF
dog	tail	How many tails does this dog have?	2	1	CF
seal	eye	How many eyes does this seal have?	4	2	CF
Japanese flag	red circle	How many red circles does this Japanese flag have?	3	1	CF



Figure 29: The 10 images used in the qualification validation set. Images with a teal border are factual (canonical count); images with a red border are counterfactual (anomalous count). See Table 6 for the corresponding questions and counts.

D.3 Annotation Protocol

Once qualified, annotators proceeded to the main evaluation. Each of the 554 counterfactual images and their factual counterparts was independently annotated by 25 qualified annotators. Rapidata displays each screen for at most 25 seconds, encouraging quick perceptual judgments rather than deliberate pixel-by-pixel analysis. This platform-imposed time limit mirrors natural human counting, which is rapid and largely automatic for small quantities. The final answer for each image was determined by majority voting over the 25 responses.

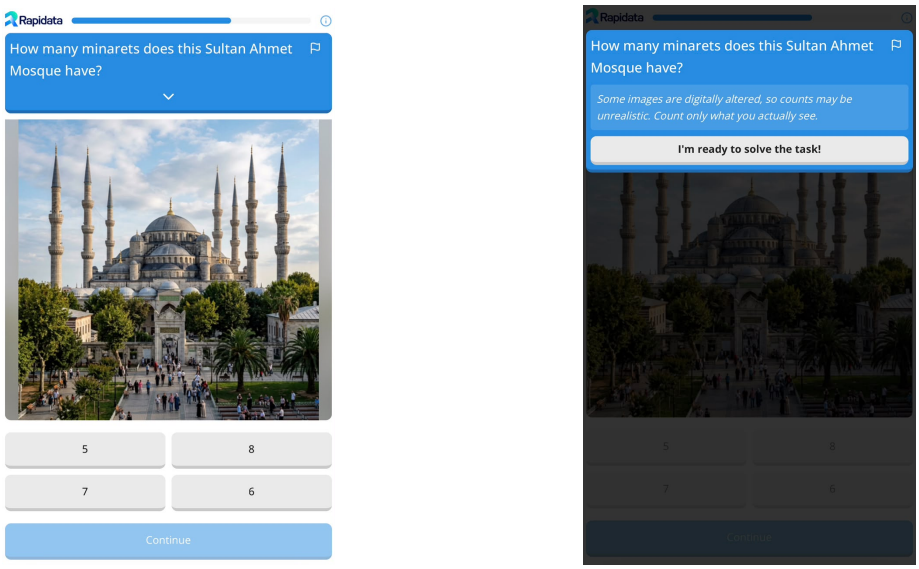
D.4 Instruction Experiment: Audience Construction

To test whether cueing annotators with the instruction *"Some images are digitally altered, so counts may be unrealistic. Count only what you actually see."* affects human performance (Section 4.1.3), we constructed two separate audiences from the *same* 10 validation images, following an identical qualification procedure. The two audiences differed in exactly one respect:

- **No-instruction audience.** Qualification and main-task screens showed only the counting question and MCQ options, with no additional instruction.
- **Instruction audience.** Both qualification and main-task screens additionally displayed the "images may be altered" caveat before the question.

Because the validation images, qualification criteria, and annotation interface were otherwise identical, any difference in performance between the two runs can be attributed to the instruction alone. Each audience independently annotated the full image set with 25 annotators per image, and final answers were again determined by majority voting.

Figure 30 shows the Rapidata annotation interface for both conditions.



(a) Without instruction.

(b) With instruction.

Figure 30: Rapidata annotation interface for the two experimental conditions. **(a)** The neutral condition shows only the counting question and four MCQ options. **(b)** The instruction condition additionally displays the caveat *"Some images are digitally altered, so counts may be unrealistic. Count only what you actually see."* above the image.