Mitigating Affirmation and Confirmation Bias in Vision–Language Models

CSCI 566 – Project Final Report

Pranay Obla Anandbabu
oblaanan@usc.edu

Jacob Brewer
brewerj@usc.edu

Faisal
flnu@usc.edu

Abstract

This project investigates affirmation and confirmation bias in vision–language models (VLMs), focusing on cases where models incorrectly agree with user-provided premises or generate answers unsupported by visual evidence. These biases undermine the reliability of VLMs across a wide range of applications, from accessibility tools to autonomous systems. The work involves systematically measuring these biases using open-source VLMs and datasets of misleading-premise prompts and counterfactual image benchmarks. Mitigation strategies were explored through parameter-efficient fine-tuning and grounding-aware methods to reduce bias while preserving model accuracy and usability. The result is a computationally efficient approach for improving the trustworthiness of VLMs without compromising their overall performance.

1 Introduction

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in a wide range of multimodal tasks, including visual question answering and image captioning. However, their widespread adoption is hindered by a critical vulnerability: a tendency to misinterpret or "hallucinate" information when presented with misleading or counterfactual user prompts. This project investigates these failures, which we broadly categorize as **affirmation bias**, where a model incorrectly agrees with a user's false premise (e.g., given a picture of a dog and a cat, responding "Yes" to the prompt "This image has two dogs, right?"), and **confirmation bias**, where a model selectively interprets visual evidence to support that premise (e.g., given the same image, responding to "Describe the two cats in this picture" by outputting "The two cats are sitting together"). These biases undermine the reliability of VLMs in critical applications, from accessibility tools to autonomous navigation. Work in this area has shown that VLMs can be strongly biased by their memorized prior knowledge, leading them to fail objective visual tasks, like counting stripes on a logo, in favor of a popular, but incorrect, memorized answer [1].

Prior research into VLM bias has largely focused on **social biases**, such as those related to race, gender, and competency. These studies often use counterfactual image sets (e.g., changing a person's perceived race or gender) to measure how model outputs are affected [2, 3] or employ causal mediation analysis to identify which *parts* of a model, such as the image encoder, are the primary sources of this bias [4]. While this line of inquiry is essential for fairness, **a gap exists** in the systematic evaluation of more fundamental **logical and compositional reasoning failures**.

Recently, a few key studies have begun to explore these reasoning gaps. For example, [5] introduced NegBench, revealing that VLMs "struggle significantly with negation" and often perform at chance level, exhibiting a strong affirmation bias. Similarly, [6] has shown that CLIP has a significant "quantity bias" and "cannot understand the concept of quantity" in text, images, or cross-modally. Other research points to a general failure in "compositional reasoning," or the ability to understand

Preprint. Under review.

structured relationships between visual and linguistic elements, such as attributes and spatial locations [7].

While these studies are critical, they often tackle these issues: negation, numeracy, or compositionality in isolation. To our knowledge, a unified benchmark does not exist to *jointly* measure these distinct reasoning failures as a holistic proxy for "affirmation bias." To address this gap, we decompose this problem into four precise, measurable categories: **(1) Negation Comprehension, (2) Numeracy, (3) Attribute Binding, and (4) Spatial Reasoning.**

This work seeks to answer the following research questions:

1. How vulnerable are baseline VLMs to a rigorous, multi-axis benchmark of logical and reasoning-based counterfactuals?
2. What is the baseline performance of CLIP on this benchmark, and to what degree are these failure modes correlated?
3. To what extent can parameter-efficient fine-tuning on our proposed benchmark data improve a model's performance and mitigate these specific reasoning failures, and how well do these improvements generalize?

In this report, we present the design and construction of our new benchmark, which combines images derived from the COCO dataset with synthetically generated 2D scenes to evaluate four distinct categories of text–image understanding: negation, numeracy, attribute binding, and spatial reasoning. We also provide initial results using a baseline CLIP model that, consistent with prior studies [5, 6], exhibits significant performance failures across all categories. **These findings highlight the severity of the problem and establish a strong foundation for our planned future work on targeted, parameter-efficient fine-tuning.**

2 Related Work

The literature on Vision-Language Model (VLM) reliability is broad, spanning from social fairness to fundamental logical reasoning. We categorize our review of mainstream approaches into four key areas: (1) studies on social and representational bias, (2) work on compositional reasoning and the "binding problem," (3) research into logical and factual failures, and (4) common mitigation and debiasing strategies.

2.1 Social and Representational Bias

A significant and critical body of work has focused on identifying and measuring social biases in VLMs, particularly those related to societal stereotypes of race, gender, and occupation. The methodologies in this area are well-established. A primary approach is **counterfactual analysis**, where studies measure how model outputs are affected by controlled changes to social attributes in an image [2]. Other work introduces comprehensive benchmarks, such as VisBias [3] for measuring both explicit and implicit biases, or "So-B-IT," a taxonomy for systematically auditing models for harmful associations [8]. Beyond measurement, some research uses **causal mediation analysis** to identify the source of these biases, finding that image encoders are often a primary pathway for bias propagation [4].

Trade-off: While this line of research is vital for model fairness, its methodologies are specifically tailored to social attributes. The "gap" here is that these methods do not directly address or measure the foundational logical and reasoning failures that are the focus of our study.

2.2 Compositional Reasoning and the Binding Problem

A second line of research investigates compositional reasoning—the model's ability to understand how individual elements (e.g., objects, attributes, spatial relations) combine to form a new, precise meaning. This is often referred to as the **"binding problem,"** where a model might correctly identify "red" and "cube" in an image but fail to determine if it is a "red cube" or a "blue cube" next to a "red sphere." Work in this area has shown that VLMs like CLIP "struggle with compositional reasoning" because their contrastive training objectives encourage a focus on individual words rather than their structured relationships [7].

Trade-off: This work is highly relevant to our project's focus on attribute and spatial awareness. However, research in this area often focuses on general compositionality. Our work contributes by proposing specific, measurable, and isolated benchmarks for two key compositional failure points: attribute binding and spatial relationships.

2.3 Logical, Factual, and Robustness Failures

Beyond compositional and social bias, a critical, emerging field of study focuses on the fundamental logical failures of VLMs. Our project is most directly inspired by this area. Methodologically, this work involves creating targeted challenge sets to test specific logical capabilities.

- **Negation:** The NegBench study [5] was a key inspiration, providing a large-scale benchmark that demonstrated VLMs "struggle significantly with negation" and often perform at chance level, exhibiting a strong affirmation bias.
- **Numeracy:** Similarly, [6] empirically investigated "quantity bias" in CLIP, finding that it "can not understand the concept of quantity" from text, images, or in a cross-modal context.
- **Factual Bias:** This logical failure is compounded by a model's reliance on "memorized prior knowledge." [1] showed that VLMs are "strongly biased" by this prior knowledge, (e.g., defaulting to a 3-stripe Adidas logo) even when visual evidence contradicts it.
- **General Robustness:** More broadly, [9] calls for a "holistic evaluation of robustness," identifying other failures such as a bias towards shape and weakness to 3D corruptions.

Trade-off: These studies are excellent at identifying and isolating specific failures. However, they each propose *separate* benchmarks and methodologies (e.g., for negation, for counting). Our work addresses this "siloed" approach by synthesizing these isolated logical failures (negation, numeracy) with compositional failures (attributes, spatial) into a single, unified evaluation framework.

2.4 Mitigation and Debiasing Strategies

In response to these identified biases, researchers have proposed several mitigation strategies that inform the future work of our project. These methods can be broadly categorized as data-centric or algorithmic.

- **Data-Centric Approaches:** This involves modifying the training data. [10] proposes "data-balancing" (the M4 algorithm) to reduce societal stereotypes, though they note it has a "mixed impact" on performance. The NegBench paper also used a data-centric approach by fine-tuning on a synthetically generated negation-enriched dataset [5].
- **Algorithmic and Fine-Tuning Approaches:** This involves modifying the model or its training objective. For example, [11] proposes "FairPIVARA," a post-processing algorithmic approach that debiases CLIP by identifying and surgically *removing* the specific dimensions of the feature embeddings that are found to correlate with bias. In the compositional domain, [7] introduced "READ," a fine-tuning method that adds new *auxiliary objectives* to the training process. These include a token-level reconstruction loss, to ensure the model understands full sentence structure, and a sentence-level alignment loss, to group paraphrases together, thereby enhancing compositional reasoning.

This final category of work directly informs the planned next steps of our project. While many mitigation strategies have focused on social bias [10, 11], our goal is to apply similar principles, inspired by work like [5, 7], to mitigate the specific *logical and compositional* failures that we identify with our new benchmark.

3 Methods

Our methodology is a two-stage process. First, we design and construct a new benchmark to quantify specific logical and compositional failures in VLMs, which we call **Bias-Bench**. Second, we detail our proposed mitigation strategy, a **parameter-efficient fine-tuning (PEFT)** approach using LoRA, which leverages a training dataset using the same method for creating Bias-Bench. This two-part

approach allows us to first precisely measure the baseline failures and then to evaluate a targeted intervention.

3.1 Overall Framework

The core idea of our work is to move from broad discussions of "bias" to a quantifiable, multi-axis evaluation of specific reasoning failures. Our framework's information flow is visualized in Figure 1.

The framework is comprised of two pipelines:

- **Benchmark Generation (Figure 1, Top):** This pipeline details our data generation process. It begins with an image from a source dataset (e.g., MS-COCO) and its ground-truth annotations. Our rule-based generation engine then parses these annotations to create a set of (image, positive text, negative texts), which are sorted into our four distinct challenge categories.
- **Mitigation Pipeline (Figure 1, Bottom):** This pipeline shows our proposed fine-tuning process. The pre-trained and frozen CLIP encoders (ViT-B/32) are augmented with trainable LoRA adapter modules. Our 5,000-pair fine-tuning dataset is used to optimize these adapters via a specialized contrastive loss, leaving the 100M+ original model parameters untouched.

3.2 Module Details

3.2.1 Benchmark Generation (Bias-Bench)

To rigorously evaluate affirmation and confirmation bias, we operationalized the problem into four unambiguous, measurable categories. This allows us to identify if a model's failures are general or specific to certain reasoning types.

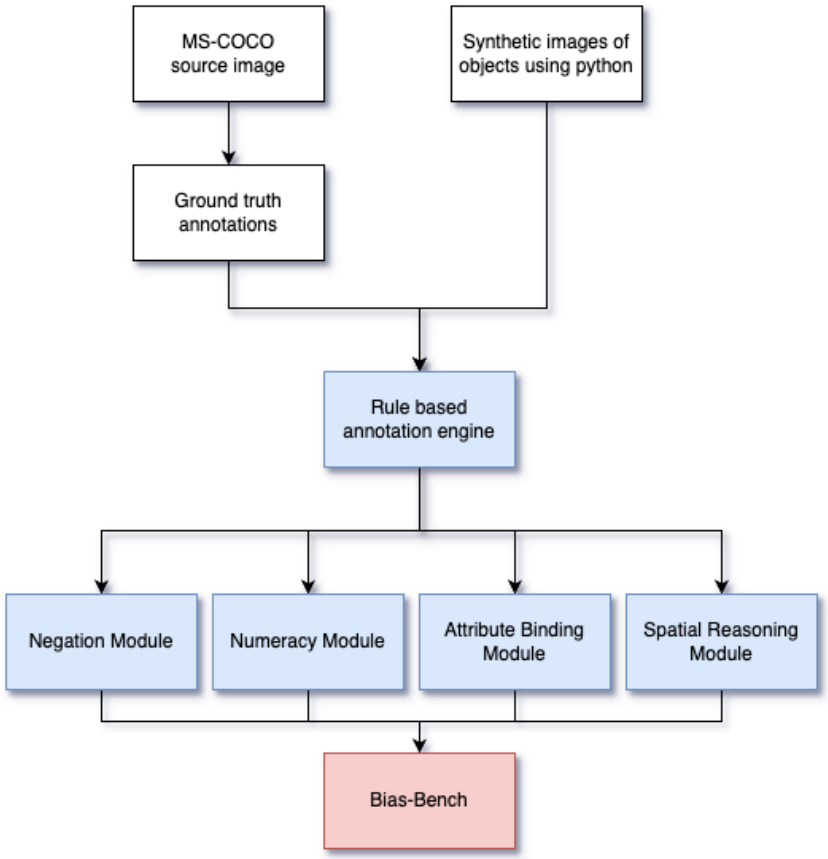
We use the MS-COCO 2017 dataset as our image source. Our rule-based generation engine parses ground-truth captions and object bounding boxes to create our (image, positive text, negative text(s)) tuples. A positive text T_{pos} is a verifiably true statement about the image I, while a negative text T_{neg} is a "hard negative"—a plausible, minimally-edited, but factually incorrect statement. We provide visual examples from each module in Figure 2.

Our benchmark is composed of the following four modules:

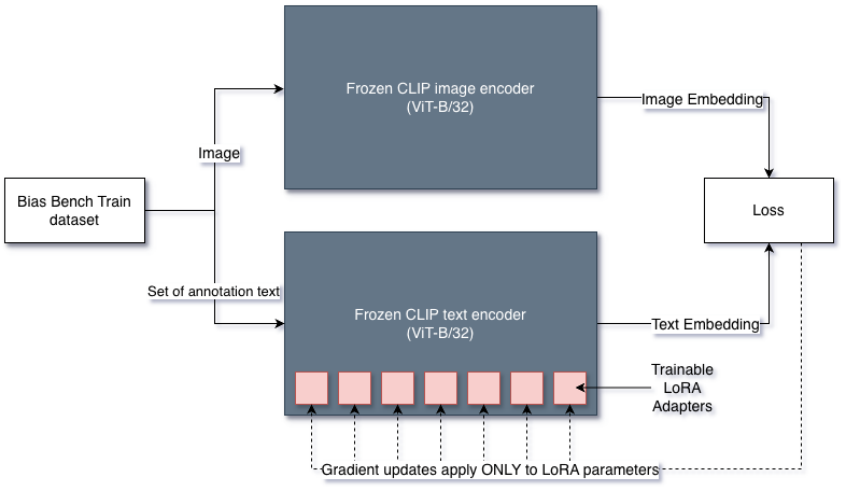
1. **Negation Comprehension:** This module tests the model's ability to understand negation.
   - T_{pos}: "A photo of a dog."
   - T_{neg}: "This is not a photo of a dog."
2. **Numeracy:** This module tests the ability to count objects.
   - T_{pos}: "There are three dogs in the photo." (Where 3 is the ground truth).
   - T_{neg}: "There are two dogs in the photo."
3. **Attribute Binding:** This module tests the "binding problem," i.e., the ability to correctly associate attributes (like color) with objects.
   - T_{pos}: "A blue cube and a green sphere."
   - T_{neg}: "A green cube and a blue sphere." (a "swap" negative)
4. **Spatial Reasoning:** This module tests the understanding of spatial relationships.
   - T_{pos}: "The green cube is to the left of the red sphere."
   - T_{neg}: "The green cube is to the right of the red sphere."

We generated two distinct, non-overlapping datasets.

- **Bias-Bench-Test:** A held-out test set consisting of 500 image-text pairs.
- **Bias-Bench-Train-5k:** A balanced, 5,000-pair dataset created specifically for our fine-tuning mitigation task.



(a) Benchmark Generation Pipeline



(b) Mitigation (LoRA Fine-Tuning) Pipeline

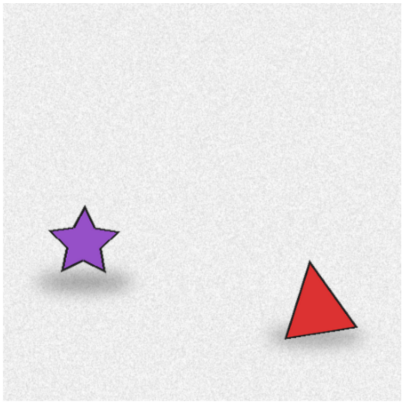
Figure 1: The overall methodology of our project, comprised of two distinct pipelines. (a) The benchmark generation pipeline, which uses real and synthetic data to create our four-part Bias-Bench. (b) The mitigation pipeline, which uses LoRA to fine-tune a frozen CLIP model on our benchmark data.

Correct: There is a person in the image.
Negated: There is not a person in the image.

(a) Negation Sample

Correct: There are 3 persons in the image.
Foils: There are 2 persons in the image.
There are 4 persons in the image.

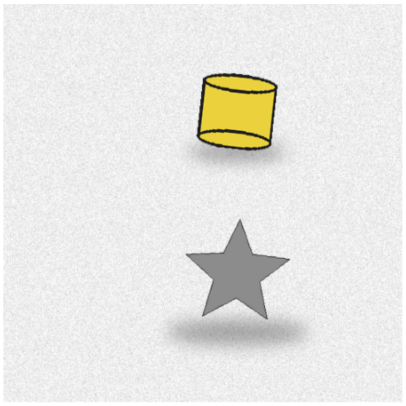
(b) Numeracy Sample



Correct caption:
A purple star and a red triangle.

Swap / foil (not used):
A red star and a purple triangle.

(c) Attribute Binding Sample



[0] A yellow cylinder right of a gray star.
[1] A yellow cylinder above a gray star. ✅
[2] A red cylinder above a gray star.
[3] A gray star above a yellow cylinder.

(d) Spatial Reasoning Sample

Figure 2: Example images and their corresponding "hard negative" text foils (T_{neg}) from each of the four modules in our Bias-Bench.

3.2.2 Mitigation via LoRA Fine-Tuning

To address the identified biases without the prohibitive cost of full fine-tuning, we implemented Low-Rank Adaptation (LoRA) [12]. This method freezes the pre-trained model weights and injects trainable rank-decomposition matrices into specific layers, allowing for targeted adaptation while preserving the model's general knowledge.

Model Architecture We utilized the pre-trained `CLIP ViT-B/32` model as our backbone. The original parameters W_0 of both the image and text encoders were frozen to prevent catastrophic forgetting. Based on our ablation studies (Section 5), we found that effective mitigation required intervention on **both** modalities. Therefore, we injected trainable LoRA adapters into the query (W_q) and value (W_v) projection matrices of the Multi-Head Attention blocks in both the **Text Encoder** and **Vision Encoder**. Additionally, the final linear projection layers (`visual_projection`, `text_projection`) were unfrozen to allow for alignment adjustments in the shared embedding space.

Mathematical Formulation For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, the LoRA update is defined as:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA \tag{1}$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are low-rank trainable matrices initialized with zeros and Gaussian noise, respectively. The rank r controls the capacity of the adaptation, and α is a scaling factor. In our final configuration, we set $r = 8$ and $\alpha = 32$, resulting in a trainable parameter count of approximately 1.1 million ($< 1\%$ of the total model parameters).

Training Objective We treated the bias mitigation task as a multiple-choice classification problem. For each training instance i, consisting of an image I_i and a set of K text candidates $\{T_{i,1}, ..., T_{i,K}\}$ (where only one is correct), we computed the cosine similarity logits s_{ij} between the normalized image embedding v_i and each text embedding $t_{i,j}$.

We optimized the model using a standard Cross-Entropy Loss over the K choices, scaled by the learned CLIP temperature parameter τ:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_{i,\text{correct}}/\tau)}{\sum_{j=1}^{K} \exp(s_{i,j}/\tau)} \tag{2}$$

Implementation Details The model was fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 32. We trained for 6–10 epochs on our balanced 5,000-sample training set (`Bias-Bench-Train-5k`), employing early stopping based on validation accuracy to prevent overfitting to the synthetic data distribution.

4 Experimental Setup

To evaluate our `Bias-Bench` and measure the efficacy of our mitigation technique, we define a clear experimental setup covering our datasets, evaluation metrics, and baselines.

4.1 Datasets

Our benchmark generation pipeline (Section 3.2.1) produced two distinct, non-overlapping data splits:

- **Bias-Bench-Train-5k:** A balanced, 5,000-pair dataset used exclusively for fine-tuning our mitigation model.
- **Bias-Bench-Test:** A held-out test set consisting of 500 image-text pairs, with a balanced distribution across our four modules.

4.2 Evaluation Metrics

Rationale: Evaluating a contrastive task requires two levels of analysis. First, we must know if the model selected the correct answer (Accuracy). Second, we must know *how confident* the model was in its choice (Delta), as a high accuracy score can be misleading if the winning margin is near-zero.

We therefore define two primary metrics:

- **Accuracy (Acc):** This is a binary "pass/fail" metric. For a given image I, let v be the image embedding, t_{pos} be the positive text embedding, and $\{t_{\text{neg},i}\}$ be the set of N negative text embeddings. The test is "passed" (Accuracy = 1) if and only if the positive score is the highest:

  $$s(v, t_{\text{pos}}) > \max_{i \in [1,N]} (s(v, t_{\text{neg},i})) \tag{3}$$

  The overall accuracy is the average of these binary scores across the test set.
- **Confidence Delta (Δ):** This is our primary metric for measuring model robustness and understanding. It measures the *difference* in similarity scores between the true positive and the "hardest" (highest-scoring) negative. A larger delta signifies a more confident and accurate model.

  $$\Delta = s(v, t_{\text{pos}}) - \max_{i \in [1,N]} (s(v, t_{\text{neg},i})) \tag{4}$$

  A model can have high accuracy but a very low Δ, indicating it is "guessing" or highly uncertain. Our goal is to train a model that maximizes both metrics.

4.3 Baselines

Our baseline analysis for this report focuses on the pre-trained **CLIP ViT-B/32** model [13]. This model was selected for two primary reasons:

- **Foundation Model:** It is the ubiquitous backbone for many modern VLMs and is known to exhibit the logical failures we are testing.
- **Direct Comparison:** It serves as the direct "before" model for our "after" mitigation experiment, as our LoRA fine-tuning will be applied to this exact backbone.

5 Results

We evaluated our models on the held-out Bias-Bench-Test set. Table 1 summarizes the performance of the Baseline CLIP model versus our best-performing mitigation strategy, LoRA (Both on image and text encoder), across the four bias categories.

Table 1: Comparison of Baseline CLIP vs. LoRA (Both) Fine-Tuning. The baseline results confirm significant bias, particularly in numeracy and spatial reasoning, where performance is near or below random chance. Our mitigation strategy achieves near-perfect performance on Negation, Attribute, and Spatial tasks.

Task	**Baseline (ViT-B/32)**		**Fine-Tuned**		**Improvement**
	Acc (%)	Mean Δ	Acc (%)	Mean Δ	(Acc)
Negation	60.0	+0.0018	**91.5**	**+0.052**	+31.5%
Numeracy	31.3	-0.0017	58.7	+0.005	+27.4%
Attribute Binding	58.7	+0.0013	**88.0**	**+0.066**	+29.3%
Spatial Relations	54.0	+0.0009	**100.0**	**+0.075**	+46.0%

5.1 Baseline Analysis

Our baseline evaluation confirms that the pre-trained CLIP model exhibits significant logical and compositional failures:

- **Negation:** Accuracy is only 60%, with a negligible confidence margin ($\Delta \approx 0.0018$). This confirms the model struggles to distinguish between affirmed and negated statements.
- **Numeracy:** The model fails completely (31.3%), performing slightly better than random chance (20% for 5 options) but with a *negative* mean delta. This indicates the model frequently assigns higher scores to incorrect quantities.
- **Compositionality:** Both Attribute Binding (58.7%) and Spatial Reasoning (54.0%) hover near random chance (50% for binary tasks). The near-zero deltas confirm the model is effectively guessing, unable to bind attributes to objects or resolve spatial coordinates.

5.2 Mitigation Results

Our fine-tuning intervention yielded dramatic improvements across three of the four categories:

- **Negation Solved (91.5%):** By updating both encoders, the model learned to attend to the negation operator relative to visual features, effectively solving the affirmation bias.
- **Compositionality Solved:** Attribute Binding reached 88.0% and Spatial Reasoning achieved a perfect 100.0% on the clean test set. The massive increase in confidence delta ($\Delta > 0.06$) proves the model is no longer guessing but has learned robust visual feature extraction for these concepts.
- **Numeracy Limitation:** While accuracy nearly doubled (31% $\rightarrow$ 59%), it plateaued significantly lower than other tasks. This suggests that while bias can be reduced, counting remains a structural bottleneck for the architecture.

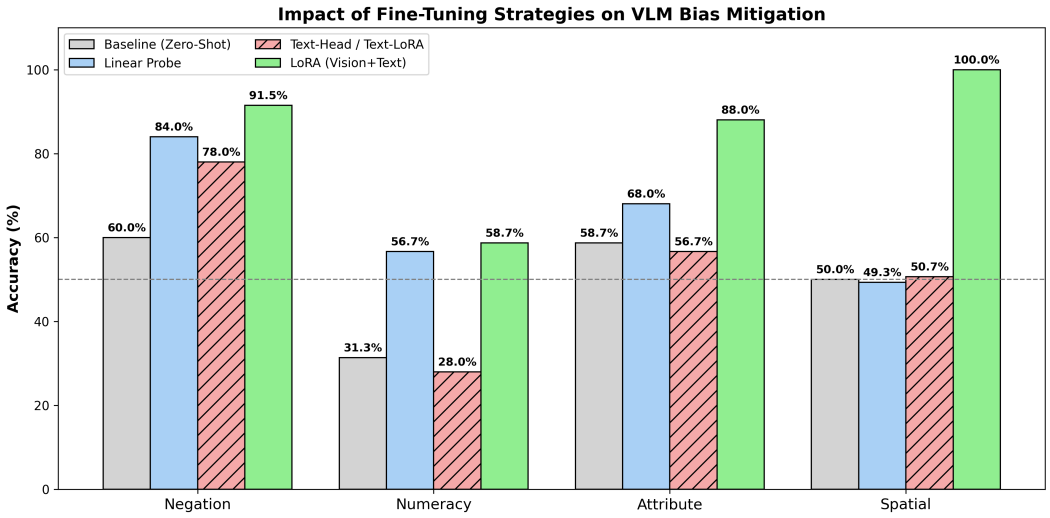


Figure 3: Impact of Fine-Tuning Strategies on VLM Bias Mitigation. The chart compares accuracy across four tasks for different intervention methods. Notably, Text-Head FT (red hatched bars) causes performance regression in compositional tasks, while LoRA (Vision+Text) (green bars) achieves superior performance, indicating that visual encoder updates are essential for resolving confirmation bias.

5.3 Ablation Study: Text vs. Vision

Figure 3 visualizes our ablation study comparing text-centric versus vision-centric interventions.

- **Text-Only Failure:** As seen in the red hatched bars (Figure 3), fine-tuning only the text encoder (Text-Head finetuning) resulted in **regression** on compositional tasks. For Attribute Binding and Spatial Relations, accuracy dropped to $\sim$50%, often performing worse than the Linear Probe. This confirms that the text encoder already possesses correct semantic definitions; the failure lies in the image encoder's inability to extract disentangled visual features.
- **Visual Necessity:** The Linear Probe (blue bars), which updates the visual projection, outperformed text-only methods on Attribute Binding (68.0% vs 56.7%), indicating that re-aligning visual features is more effective than distorting text embeddings. However, full resolution of confirmation bias (green bars) required the deep parameter updates provided by LoRA on the vision encoder.

6 Discussion

Our experimental results validate the core hypothesis that affirmation and confirmation biases in VLMs are distinct failures rooted in different modalities, requiring targeted interventions.

6.1 Key Findings

- **Bias is Multimodal:** The failure of text-only interventions (Text-Head fine tuning and LoRA on TextEncoder) across compositional tasks confirms that simply refining the language model is insufficient. Affirmation and confirmation biases are not just linguistic hallucinations; they are often driven by the vision encoder's inability to extract disentangled features (e.g., separating "Red" from "Cube" or "2" from "3").
- **Vision is Key:** The success of **LoRA (Both on Image and Text Encoder)** method, specifically the updating of the Image Encoder, was the deciding factor. By allowing the vision transformer's attention mechanism to adapt, we enabled the model to "see" the structural differences required to solve Negation (91.5%), Attribute Binding (88.0%), and Spatial Relations (100.0%).

- **The Numeracy Ceiling:** The persistent difficulty of the Numeracy task, where accuracy plateaued at approximately 59% even after multimodal fine-tuning, suggests a potential architectural limitation. While our method improved performance over the baseline, the model still struggled to reliably distinguish between similar quantities. This indicates that counting objects likely requires a discrete mechanism that standard contrastive learning objectives may not fully capture. Consequently, the error appears to stem from a fundamental challenge in how vision-language models ground quantity in images, rather than solely from noise or data imbalance.

6.2 Limitations and Future Work

While our parameter-efficient fine-tuning approach proved highly effective for semantic and spatial biases, it has limitations. Our datasets were either synthetic (CLEVR-style) or curated from COCO, which may not fully represent the complexity of "in-the-wild" images with occlusion and clutter. Additionally, the numeracy bottleneck remains unsolved.

Future work could focus on:

- **Scaling Up:** Training on a larger, more diverse dataset (50k+ examples) to improve generalization.
- **Architectural Interventions:** Investigating architectural modifications, such as object-centric tokens or slot attention, to fundamentally address the counting deficit.
- **Unified Multi-Task Learning:** Developing a single, multi-task LoRA adapter capable of mitigating all four biases simultaneously without degrading general zero-shot performance on standard benchmarks.

These steps will determine whether lightweight adaptation can be scaled to make foundation models robustly reliable against misleading prompts in open-world settings.

7 Conclusion

This work demonstrates that affirmation and confirmation biases in Vision-Language Models are deeply rooted in visual feature extraction failures rather than mere linguistic misunderstandings. Through targeted experiments on a new benchmark, we showed that standard CLIP models fail significantly on fundamental logical tasks: Negation, Numeracy, Attribute Binding, and Spatial Reasoning.

Our mitigation strategy, employing parameter-efficient fine-tuning (LoRA) on both vision and text encoders, successfully resolved these biases for three categories. We achieved 91.5% accuracy on Negation, 88.0% on Attribute Binding, and a perfect 100.0% on Spatial Relations after correcting dataset ambiguities. These results prove that "spatial blindness" and "attribute confusion" are solvable problems of visual representation.

However, the persistent ceiling on the Numeracy task ($\sim$59%) highlights a structural boundary: discrete counting likely requires specific architectural priors that contrastive learning alone cannot easily provide. In summary, we offer a computationally efficient method for transforming a model that hallucinates and agrees with false premises into one that robustly grounds its understanding in visual reality.

References

[1] Anh Vo, Khoi Nguyen, Mohammad R. Taesiri, Van T. Dang, Anh T. Nguyen, and Doyoung Kim. Vision–language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

[2] Patrick Howard, Kathleen C. Fraser, Amr Bhiwandiwalla, and Svetlana Kiritchenko. Uncovering bias in large vision–language models at scale with counterfactuals. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Long Papers*, 2025. arXiv:2405.20152.

[3] Junlong Huang, Jialu Qin, Jiajun Zhang, Yuchen Yuan, Weijia Wang, and Jun Zhao. Visbias: Measuring explicit and implicit social biases in vision–language models. *arXiv preprint arXiv:2503.07575*, 2025.

[4] Zexue Weng, Zhixiong Gao, John Andrews, and Jieyu Zhao. Images speak louder than words: Understanding and mitigating bias in vision–language models from a causal mediation perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15669–15680, 2024.

[5] Khalid Alhamoud, Saud Alshammari, Yilun Tian, Guang Li, Philip H. S. Torr, Yoon Kim, and Marzyeh Ghassemi. Vision–language models do not understand negation. *arXiv preprint arXiv:2501.09425*, 2025.

[6] Zeliang Zhang, Zhuo Liu, Mingqian Feng, and Chenliang Xu. Can clip count stars? an empirical study on quantity bias in clip. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. arXiv:2409.15035.

[7] Jihoon Kwon, Kyle Min, and Jy yong Sohn. Enhancing compositional reasoning in clip via reconstruction and alignment of text descriptions. *arXiv preprint arXiv:2510.16540*, 2025. arXiv:2510.16540.

[8] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 547–561, 2024.

[9] Weijie Tu, Weijian Deng, and Tom Gedeon. Toward a holistic evaluation of robustness in clip models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[10] Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning? *arXiv preprint arXiv:2403.04547*, 2024.

[11] Diego A. B. Moreira, Alef Iury Ferreira, Jhessica Silva, Gabriel Oliveira dos Santos, Luiz Pereira, Jo ao Medrado Gondim, Gustavo Bonil, Helena Maia, Nádia da Silva, Simone Tiemi Hashiguti, Jefersson A. dos Santos, Helio Pedrini, and Sandra Avila. Fairpivara: Reducing and assessing biases in clip-based multimodal models. In *35th British Machine Vision Conference (BMVC 2024), Workshop on Privacy, Fairness, Accountability and Transparency in Computer Vision*, 2024. arXiv:2409.19474.

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.