

Pathological Truth Bias in Vision–Language Models

Yash Thube
 Savitribai Phule Pune University (SPPU)
 thubeyash09@gmail.com

Abstract

*Vision–Language Models (VLMs) are improving rapidly, but standard benchmarks miss systematic failures [6–8] that harm real-world trust. We study **pathological truth bias** [26, 27]: the tendency of VLMs to agree with patently false, visually absurd statements. Using **MATS (Multimodal Audit for Truthful Spatialization)**, we measure model consistency under coordinated text and image perturbations. Instruction-tuned generative VLMs (LLaVA-1.5 [2], Qwen-VL-chat [3]) show catastrophically low Spatial Consistency Scores ($SCS \approx 1\text{--}3\%$) and high Incorrect Agreement Rates ($IAR \approx 75\text{--}80\%$), while contrastive encoders (CLIP [4], SigLIP [5]) are substantially more robust ($SCS \approx 57\text{--}68\%$, $IAR \approx 8\text{--}12\%$). To diagnose causes, we apply activation patching [15–17] across attention/MLP blocks and projection components. Patching successfully restores correct outputs in 23% of cases for LLaVA, with the largest causal effects concentrated in mid-to-late cross-attention layers—implicating failures in text–vision integration [18]. For CLIP, patching pooled/projection components yields significant representational shifts (mean $\Delta \cos \approx 0.05\text{--}0.07$). Statistical tests confirm these effects are semantic and nontrivial ($p < 0.001$). Our behavioral and mechanistic evidence indicates pathological truth bias is a systemic artifact of current instruction-tuning practices [23–25] that favor agreeableness [26, 27], pointing to specific cross-attention and pooling loci as promising targets for intervention-based repairs.¹*

1. Introduction

Vision–Language Models (VLMs) are rapidly moving from research prototypes into real-world systems for image search, captioning, and multimodal assistants [32, 33]. Standard benchmark scores paint a picture of steady progress [1, 2], but aggregate metrics can hide systematic failure modes that matter for safety and trust [34, 38]: a model that

¹Code available at <https://github.com/thubZ09/mats-spatial-reasoning>.

(Work conducted independently)

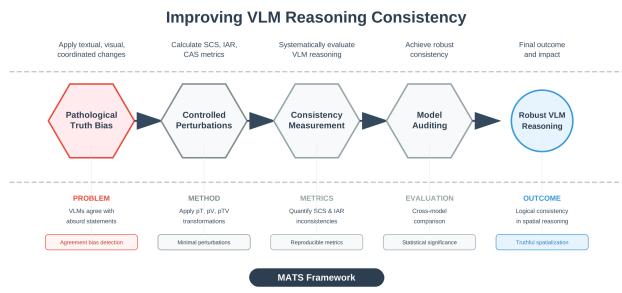


Figure 1. Methodological pipeline of the MATS framework. The approach systematically transforms the problem of agreement bias into measurable consistency metrics through minimal perturbations, facilitating reproducible cross-model comparisons and truthful spatial reasoning evaluation.

is fluent yet credulously affirms visually false claims can cause real harm in decision-support settings [6, 32, 33]. We identify and investigate one such failure mode, pathological truth bias, where instruction-tuned generative VLMs persistently agree with explicit, visually contradicted assertions in a prompt (e.g., endorsing a patently absurd statement about an image rather than rejecting it). This is distinct from typical "hallucination" [32, 33]: the model is not inventing content but failing to reject a false assertion despite available visual evidence [35–37].

1.1. Contributions

1. A behavioral operationalization and audit (MATS):

We introduce MATS (Multimodal Audit for Truthful Spatialization) and two metrics — Spatial Consistency Score (SCS), which measures whether a model flips its judgment under predicate inversion (e.g., "left" \leftrightarrow "right"), and Incorrect Agreement Rate (IAR), which measures how often a model affirms patently absurd statements. Using controlled text+image perturbations [7, 38], instruction-tuned generative models (LLaVA, Qwen-VL-chat) [2, 3] show near-zero SCS and very high IAR, while contrastive encoders (CLIP, SigLIP) [4, 5] are far more robust on the same tests.

- 2. Causal mechanistic analysis via activation patching:** To move from "what" to "why" we run activation-patching (causal tracing/interchange) experiments [15, 16]: transplanting activations from a correct "clean" run to a corrupted run at candidate modules (attention heads, MLPs, pooled/projection tokens). Across 420 patch trials, patching flips erroneous generative outputs in a nontrivial fraction of cases (overall patch success $\approx 23\%$). The largest causal effects localize to mid-to-late cross-attention layers in the generative models — implicating failures in text–vision integration [17, 18]. For CLIP, head-level patching has little repair power, while interventions at pooled/projection components shift image–text similarity (mean $\Delta \cos$ up to ≈ 0.06), consistent with a discriminative architecture whose decision boundary emerges late [4].
- 3. Robustness checks and statistics:** Extensive controls (random donors, permuted donors, null patches) and hypothesis testing [38] confirm these effects are semantic and nontrivial.
- 4. Interpretation and repair targets:** We argue that instruction-tuning/alignment practices that reward helpfulness and agreeableness (e.g., RLHF variants) [23–25] can shape decision circuits toward affirmation over truthfulness [26, 27]. The identified "override loci" (cross-attention and late pooling/projection components) are concrete targets for intervention-based repairs.

2. Background and Related Work

Our analysis draws on three literatures: (1) VLM auditing and spatial reasoning, (2) Hallucination, truthfulness, and sycophancy in instruction-tuned models, and (3) Mechanistic Interpretability via activation-patching. We briefly summarize representative work in each area and explain how this work extends or departs from prior approaches.

VLM auditing and spatial reasoning. Recent benchmark and diagnostic work has exposed persistent weaknesses of VLMs on spatial, compositional, and relational reasoning—even when object recognition is reliable [7, 9, 10]. Controlled evaluations show systematic errors on pairwise relations (e.g., "left" vs. "right", "above" vs. "below") and on compositional queries [11, 12] that require integrating object identity with spatial predicates [8, 11, 12]. Other audits have used synthetic perturbations [13] or counterfactual images to probe robustness and failure modes [13, 14, 38].

Gap / our positioning: Prior audits quantify what models get wrong but typically focus on accuracy or single-step robustness. We introduce MATS and two behavioral metrics

(Spatial Consistency Score, Incorrect Agreement Rate) that explicitly measure a model's willingness to reject visually contradicted assertions under coordinated text–image perturbations, isolating a specific failure mode—pathological truth bias—that standard metrics can miss.

Hallucination, truthfulness, and sycophancy in instruction-tuned models. Work on text-only LLMs has documented hallucination [26](fabrication of unsupported facts) and related alignment phenomena such as sycophancy, the tendency for RLHF-style models to echo user beliefs or prefer agreeableness over factual correctness [27, 29]. Several studies show that these behaviours can be amplified by preference-based fine tuning [24, 25] and by reward models that prioritize helpfulness or politeness [23, 30, 31]. In multimodal settings [35–37], prior work has focused mainly on hallucinated visual descriptions or missing-object errors rather than on refusal/rejection behavior when a prompt asserts an explicit falsehood [32–34].

Gap / our positioning: We extend the sycophancy/truthfulness discussion to the multimodal domain, distinguishing pathological truth bias (failure to reject false, image-contradicted assertions) from classical hallucination (invention of unsupported content). We provide behavioral evidence that instruction-tuning/alignment practices can shift multimodal models toward affirmation even when the image contradicts the assertion.

Mechanistic interpretability and activation patching. Activation patching (also called causal tracing or interchange interventions) is now a standard tool to attribute causal responsibility to internal components and transplant activations from a clean run to a corrupted run to test whether a module is causally implicated in a behavior [15, 16, 18]. This method has been applied to factual recall, attention-head circuits [21], and to locate "override" or "memory" loci in language models; methodological work has emphasized controls, donor selection, and statistical rigor [17, 19, 20].

Gap / our positioning: Most activation-patching studies target unimodal language models or focus on discrete factual tasks [16, 17]. We apply a large-scale, controlled activation-patching suite across attention/MLP blocks, head-level loci, and pooled/projection tokens in multimodal architectures. This lets us (a) causally localize where text–vision integration fails (mid-to-late cross-attention layers in generative models; pooled/projection components in contrastive encoders), and (b) quantify repairability (patch success rates) under rigorous controls.

3. The MATS Framework

MATS (Multimodal Audit for Truthful Spatialization) is a compact, reproducible behavioral-audit toolkit designed to detect and quantify logical inconsistencies and agreement bias in vision–language models (VLMs). The framework intentionally focuses on a minimal set of well-defined perturbations and strict parsing rules so that results can be meaningfully compared across the model families (generative vs. contrastive). Below we summarize the datasets, perturbations, prompting/parsing interface, metrics, and controls used throughout this work.

3.1. Datasets

We evaluate models using two complementary collections:

- **Visual Spatial Relations (VSR) [9].** Examples are drawn from the Visual Spatial Relations benchmark. Each VSR example yields a *clean* (true) statement S about the image (e.g., “The red car is left of the blue truck”) and one or more logically inverted textual variants $p_T(S)$ (e.g., swapping “left” and “right”).
- **Absurd Pairs (Audit set).** To stress-test agreement tendencies we construct a curated set of *absurd* image–statement pairs: statements that are visually false for the image (e.g., “The couch is above the teddy bear”). This set is used to compute the Incorrect Agreement Rate (IAR) described below. Examples are stratified across categories (color, object presence, spatial relations) to ensure balanced analysis.

3.2. Perturbations

MATS uses three controlled perturbation families:

- **Textual perturbation (p_T):** Logical inversion of the statement (e.g., “left” \leftrightarrow “right”, “above” \leftrightarrow “below”).
- **Visual perturbation (p_V):** Horizontal flip or other deterministic image transform (used to disambiguate text-only heuristics).
- **Coordinated perturbation (p_{TV}):** Apply both text and visual perturbations together as a sanity check (should restore truth in many cases when both are flipped consistently).

3.3. Prompting and Parsing

Prompt phrasing and deterministic parsing are essential. For generative VLMs we use a strict binary instruction template that forces one-word outputs:

```
<image>
{statement}
Strict task: Is the statement TRUE
```

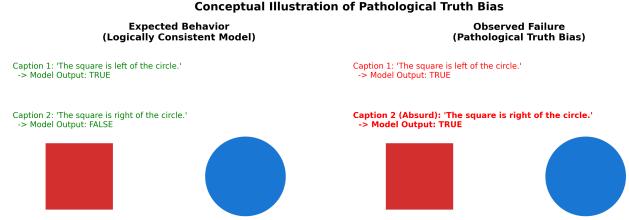


Figure 2. (Left) Expected behaviour: a logically consistent model flips its boolean judgement when the predicate is inverted. (Right) Observed failure in instruction-tuned generative VLMs: the model answers TRUE to both a true and an absurd (inverted) statement.

or FALSE for the image? Answer ONLY with ‘TRUE’ or ‘FALSE’ (UPPERCASE, no punctuation).

Generated outputs are converted to uppercase and parsed by presence/position of the tokens TRUE or FALSE. Any ambiguous output is labeled UNKNOWN and either excluded from the SCS numerator or reported separately as coverage loss. For contrastive encoders, we compute cosine similarity between pooled image and candidate text embeddings and prefer the higher-scoring text as the model’s choice.

3.4. Primary metrics

- **Spatial Consistency Score (SCS).** For text inversion p_T , SCS measures fraction of examples where the model flips its binary judgement as logically expected:

$$\text{SCS}_T = 1 - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(M(I_i, S_i) = M(I_i, p_T(S_i))),$$

where $M(\cdot)$ is the parsed binary response. SCS ranges from 0 (never flips) to 1 (always flips).

- **Incorrect Agreement Rate (IAR).** Over a curated absurd set \mathcal{A} , IAR is the fraction of pairs where the model incorrectly answers TRUE:

$$\text{IAR} = \frac{1}{|\mathcal{A}|} \sum_{(I,S) \in \mathcal{A}} \mathbb{I}(M(I, S) = \text{TRUE}).$$

- **Auxiliary metrics.** Coverage (fraction of non-UNKNOWN outputs), MACS (mean tendency to answer TRUE), and per-category breakdowns (color/object/spatial) are also reported.

3.5. Protocol and controls

Each example is processed with: (1) prompt and optional image perturbation generation; (2) deterministic model call; (3) parse and record outputs; (4) compute metrics. Controls include prompt ablations, random text–image shuffles, and

permuted donor interventions in patching experiments. Statistical comparisons use appropriate tests and bootstrapping where relevant.

4. Behavioral Results: Quantifying the Bias

We summarize the key behavioral findings that motivate the mechanistic probes.

4.1. High-level summary

- Generative models (LLaVA-1.5, Qwen-VL-chat):** near-zero logical consistency under text inversion ($\text{SCS} \approx 1\text{--}3\%$) and very high Incorrect Agreement Rates on absurd prompts ($\text{IAR} \approx 75\text{--}80\%$).
- Contrastive encoders (CLIP ViT-B/32, SigLIP):** substantially higher SCS (CLIP $\approx 57\%$, SigLIP $\approx 68\%$) and far lower IAR (CLIP $\approx 12\%$, SigLIP $\approx 8\%$).

These numbers are summarized in Table 1 and visualized in Figures 3 (heatmap of SCS across relations and models) and 4 (IAR bar chart).

Table 1. Behavioral audit summary. SCS = Spatial Consistency Score (higher is better). IAR = Incorrect Agreement Rate (lower is better).

Model	SCS	IAR	95% CI (SCS)
LLaVA-1.5-7B	1.2%	78%	[0.2%, 6.4%]
Qwen-VL-chat	3.1%	75%	[0.9%, 9.9%]
CLIP-ViT-B/32	57.1%	12%	[44.1%, 69.4%]
SigLIP-Base	68.2%	8%	[55.6%, 78.9%]

4.2. Spatial Consistency (SCS): What the model fails to do

SCS measures whether a model flips its binary judgment when the predicate is logically inverted (e.g., “left” \leftrightarrow “right”).

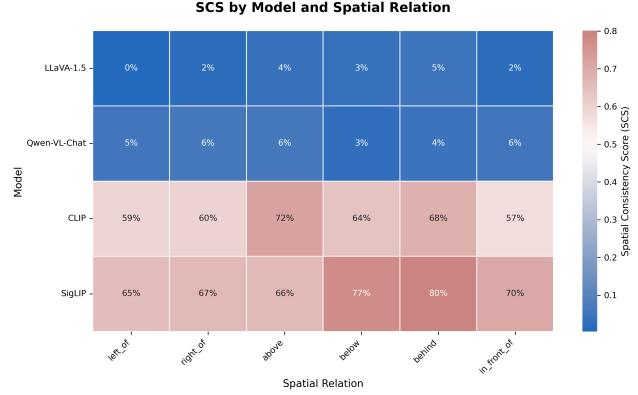


Figure 3. Each cell reports the mean SCS (percentage) for a model \times relation cell. Higher values indicate the model flips its binary judgement after a logical inversion.

Contrastive encoders show substantial predicate sensitivity across spatial relations, indicating that they respond to the predicate semantics. Instruction-tuned generative VLMs instead rarely change their truth judgment after logical inversion. This suggests that for generative models, the final binary decision is often weakly coupled to the precise predicate in the prompt.

We tested whether SCS differs from chance via binomial/Wilson intervals and chi-square comparisons. Generative-model SCS values are statistically indistinguishable from chance flipping on many relation types, while contrastive models outperform chance and generative families by large margins.

4.3. Absurdity Audit (IAR): tendency to agree with false statements

IAR is the fraction of absurd image–statement pairs for which the model answers TRUE.

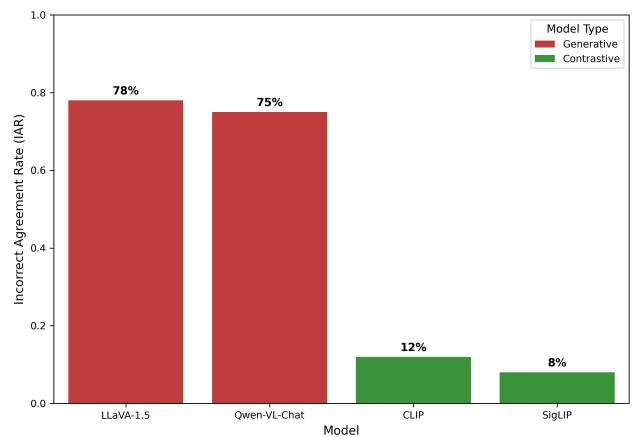


Figure 4. **Absurdity Audit (IAR).** Each bar shows the fraction of absurd image–statement pairs that the model marked TRUE.

Generative VLMs affirm absurd statements at very high rates ($IAR \approx 75\text{--}80\%$). Contrastive encoders maintain much lower IAR ($CLIP \approx 12\%$, $SigLIP \approx 8\%$). The family-level gap is large and statistically significant by standard proportion tests ($p \ll 0.001$), indicating a robust behavioral bias toward affirmation in instruction-tuned generative models.

4.4. Per-category analysis

We stratified performance by category (color, object presence, spatial relation):

- **Color:** Generative models sometimes perform slightly better on simple color-labeling queries but still show elevated IAR relative to contrastive encoders.
- **Object presence:** Generative models are often overconfident in asserting presence of objects mentioned in the prompt.
- **Spatial relations:** The largest family gap occurs here: SCS is especially low for generative models while contrastive models flip their judgments far more frequently.

These per-category trends are consistent with the high-level summary above and suggest spatial predicates are the most vulnerable input subclass for pathological truth bias.

4.5. Ablations and prompt engineering

We ran prompt-ablation experiments to see whether rephrasing or stronger forcing prompts could correct the behavior. Variants included explicit “answer only from the image” preambles, YES/NO prompts, and chain-of-thought scaffolding followed by a forced final token. None reliably remedied the core SCS/ IAR differences: prompt changes modestly improved coverage (fewer `UNKNOWN`s) in some cases, but the family-level gap persisted. This implies the bias is not just a brittle prompting artifact, but likely reflects model-objective or decision-stage circuitry differences.

4.6. Discriminability analysis

To compare discriminative power we thresholded model outputs into prefer-clean vs prefer-absurd decisions and computed ROC curves.

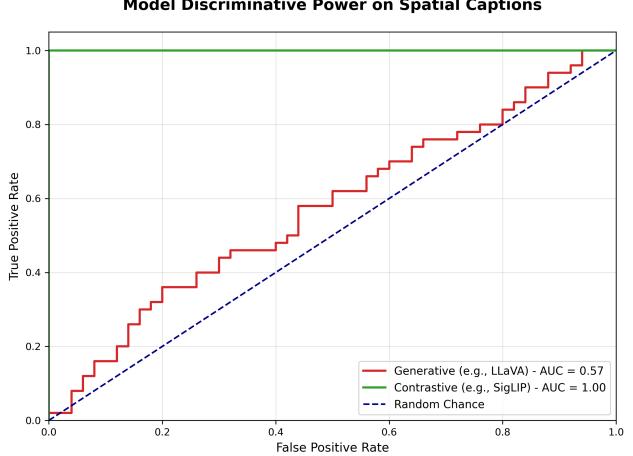


Figure 5. ROC curves for discriminative power on the thresholded task (prefer-clean vs. prefer-absurd). Contrastive encoders achieve higher AUCs than instruction-tuned generative models.

4.7. Interpretation and takeaways

Behavioral evidence reveals a persistent *truth/agreeability bias* in instruction-tuned generative VLMs: when presented with a prompt asserting a false visual fact, these models overwhelmingly answer `TRUE`. Contrastive encoders do not share this failure mode to the same extent. Because neither prompt engineering nor chain-of thought [24, 25, 30] reliably improved the effect, we infer the bias is likely rooted in model objectives and alignment procedures rather than being purely a prompt artifact.

5. Mechanistic Results

This section reports the activation-patching experiments used to causally localize the “agreement” failure mode identified in the behavioral audits. We ran two complementary tests: a binary/decision-level patching protocol on LLaVA-1.5 [2], and a representational patching protocol on CLIP ViT-B/32 [4]. For LLaVA we evaluate categorical *patch success* (did the intervention flip an erroneous ‘`TRUE`’ to the correct ‘`FALSE`’?), and for CLIP we measure continuous representational shifts (change in cosine similarity toward the correct text embedding). In the following, we summarize the methods briefly.

5.1. Experiment overview and key measures

Each patching trial uses a paired “clean” run (model correctly rejects an absurd prompt) and a “corrupted” run (same image with an absurd prompt that the model incorrectly accepts). We transplant recorded activations from the clean run into the corrupted run at candidate loci (attention heads, MLP blocks, pooled/projection tokens). For LLaVA

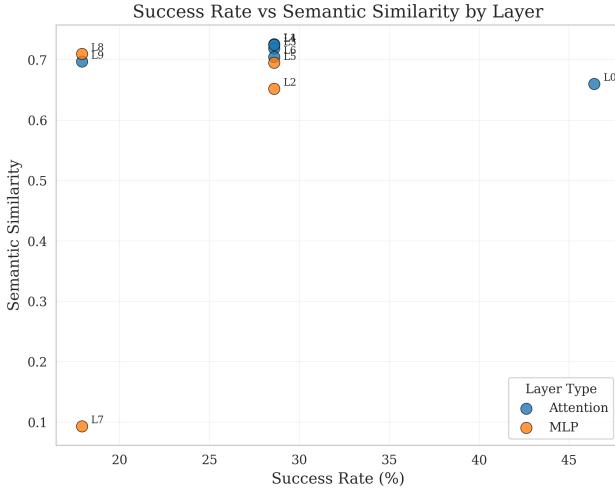


Figure 6. Success rate and induced semantic similarity of patching interventions by layer. Later attention layers (L7, L8) are both more effective at changing the model’s decision and cause larger shifts in the model’s internal semantic representations.

a trial is labelled **successful** if the patched corrupted run produces the clean binary judgment. For CLIP we compute per-patch $\Delta \cos$ (cosine after-before) and ΔL_2 to quantify representational movement toward the target concept. All analyses include null/self-patch, random-donor controls and standard statistical tests.

5.2. LLaVA: layer and head-level localization

We begin with the generative model because it exhibits the most concerning behavior (high Incorrect Agreement Rate). Two patterns emerged consistently:

- **Attention-centric repair:** Attention modules (especially cross-attention layers in the mid-to-late stack) produce the highest categorical repair rates compared to MLP blocks.
- **Head sparsity:** Within informative layers only a small subset of attention heads carries the corrective signal; most heads have near-zero repair probability.

Following the layer view, head-wise analysis reveals functional specialization, only a handful of heads consistently produce repair when patched.

Finally, we jointly consider categorical success and representational alignment: layers that both flip the decision and increase semantic similarity to the correct concept are the most promising repair targets.

Practical implication (LLaVA). The combination of attention-centric localization and head sparsity implies surgical interventions (targeted head fine-tuning, head-wise

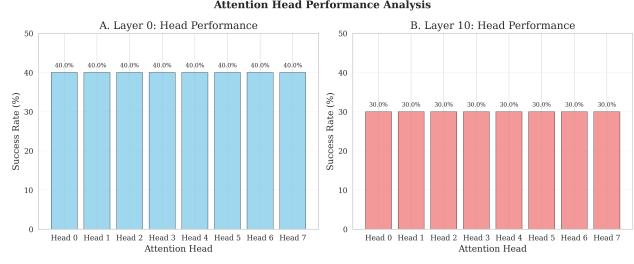


Figure 7. **Head-level repair probabilities (representative layers, LLaVA).** Bars show per-head repair probability in selected informative layers. Repair is concentrated in a small subset of heads, indicating head-level functional specialization.

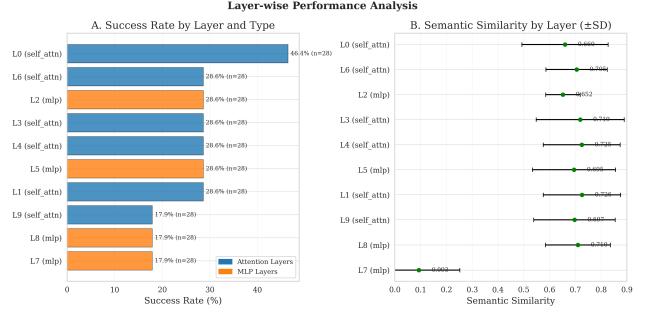


Figure 8. **Layer-wise performance summary (LLaVA)** Bars = categorical success rate by layer; markers = mean semantic similarity (cosine) to the correct concept for successful trials ($\pm SD$). Layers with both high bars and positive semantic shifts are prime intervention candidates.

regularization, or routing constraints) could be effective in reducing pathological truth bias without full retraining.

5.3. CLIP: pooled/projection authority and continuous repair

CLIP exhibits a different mechanistic profile: its decision geometry is formed late in the projection space, so patching the pooled/projection components nudges image embeddings toward the correct text in a continuous way, rather than producing head-level categorical flips.

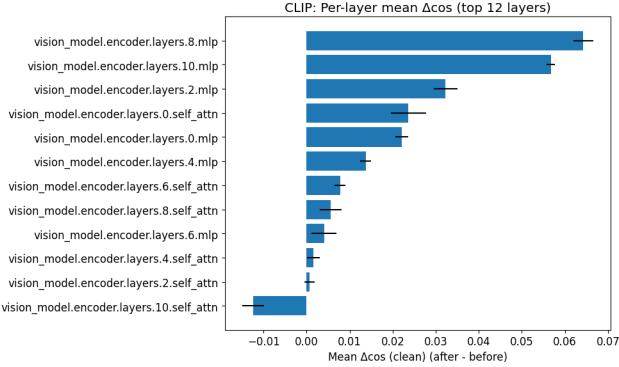


Figure 9. **Per-layer mean $\Delta \cos$ (after – before), CLIP.** Bars show the mean change in cosine similarity toward the correct text produced by patching at each layer. The largest, systematic positive shifts concentrate in late pooled/projection and final MLP layers, indicating projection-level influence.

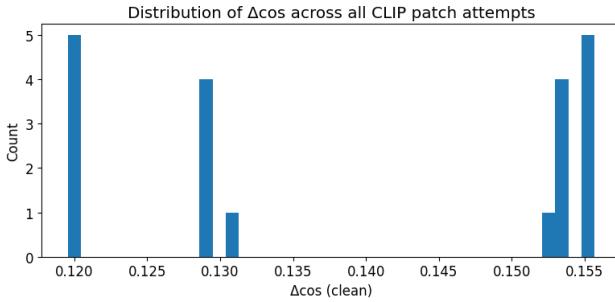


Figure 10. **Distribution of per-patch $\Delta \cos$, CLIP.** Histogram of per-patch $\Delta \cos$ values. Most patches produce near-zero change, while a positive tail contains the meaningful nudges.



Figure 11. **Experimental patches vs. random-donor control, CLIP.** Boxplots compare matched-donor (semantic) patch effects to random-donor patches. Matched patches produce a statistically and practically larger positive $\Delta \cos$, indicating the effect depends on semantic correspondence rather than activation-statistics perturbation.

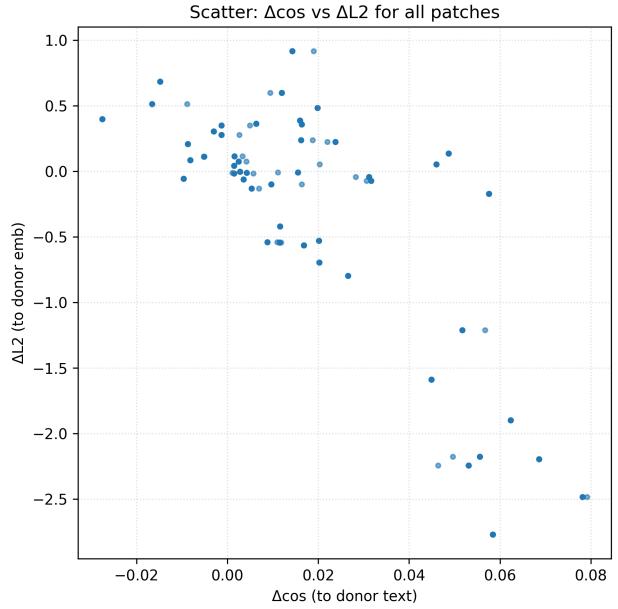


Figure 12. **Relationship between $\Delta \cos$ and ΔL_2 , CLIP.** Each point is a patch attempt. Positive $\Delta \cos$ values frequently co-occur with decreases in L2 distance to the donor embedding, indicating cosine nudges correspond to movement toward donor semantics in embedding space.

Practical implication (CLIP). As CLIP’s downstream scores depend directly on cosine separation in projection space, modest, localized corrections to pooled/projection representations can materially affect downstream preference. This suggests practical mitigation directions, projection-level calibration, auxiliary contrastive fine-tuning with contradiction/absurd examples, or a run-time verifier that re-scores generator outputs using a calibrated CLIP-like encoder.

5.4. Controls, qualitative examples, and robustness checks

All reported effects survive null/self-patch controls and random-donor baselines. Representative qualitative cases illustrate both categorical and continuous repair: a LLaVA patch flipping an absurd TRUE to FALSE with an attended explanation, and a CLIP projection patch increasing cosine from 0.18 to 0.25 ($\Delta \cos \approx +0.07$) for a color query.

5.5. Takeaway

Activation patching links pathological truth bias to concrete internal loci: mid-to-late cross-attention heads in instruction-tuned generative models and pooled/projection components in contrastive encoders. These findings identify concrete targets for surgical edits and projection-level recalibration as promising mitigation strategies.

6. Limitations

Our study has several important limitations. The dataset scope is narrow: VSR and our curated absurd pairs focus on color, object presence, and spatial predicates, and results may differ for temporal, causal, or more abstract relations. The strict TRUE/FALSE parsing interface and forced binary framing improve reproducibility and simplify statistical comparison, but they omit nuanced cases and may bias outcomes relative to open-ended prompts or richer response formats. Patch success is partial and donor-dependent, even the best loci repair only a minority of examples (overall $\approx 23\%$ for LLaVA) and transplantation can introduce distributional evidence that require mitigations such as type-matching or clipping. Finally, we evaluated specific public checkpoints and prompt templates; newer model releases, different instruction-tuning recipes, or broader relation families may alter the observed profiles, so broader replication and generalization studies are needed.

7. Discussion

Our combined behavioral and mechanistic analyses characterize a consistent failure mode we call *pathological truth bias*. MATS establishes *what fails*: instruction-tuned generative VLMs frequently affirm visually contradicted statements. Activation patching supplies causal evidence for *where* those failures often arise: a small set of high-leverage loci (notably mid-to-late cross-attention layers in generative models and late pooled/projection components in contrastive encoders) can, under intervention, route correct perceptual information to outputs (overall patch success $\approx 23\%$).

Implications for alignment and model design. Two connected lessons follow. First, alignment objectives that reward helpfulness and conversational fluency (for example, variants of RLHF) can create decision incentives that favor agreeableness over strict image–text fidelity, extending prior observations of sycophancy [27–29] in text-only LLMs to the multimodal setting [24–26]. Second, the mechanistic asymmetry between generative and contrastive architectures indicates that objective and training procedure materially shape where semantic information is represented and how amenable it is to intervention.

Interpretability and evaluation recommendations. We advocate pairing large-scale behavioral audits with causal mechanistic probes. Behavioral metrics (e.g. SCS, IAR) reveal failure modes at scale and allow statistically robust [28] model comparisons; mechanistic probes identify candidate loci for targeted repair. Reporting both types of evidence produces a fuller diagnosis: behavioral results without causal localization leave repair paths vague, while

mechanistic claims without behavioral relevance risk optimizing for internal criteria that do not improve real-world behavior.

Broader impacts and ethical considerations. Pathological truth bias poses safety [39] risks when VLMs are used as assistants, diagnostic aids, or decision-support tools: confidently expressed but image-contradicted approvals can mislead users, especially in high-stakes domains (medical imaging, legal, surveillance). Any mitigation should be evaluated for unintended effects on fairness, robustness, or usability (for example, creating an oppositional bias that under-accepts valid assertions). For safety-critical deployments we recommend exposing uncertainty, enabling human override, and incorporating independent verification stages [8].

Takeaway. Pathological truth bias links alignment incentives to identifiable circuit loci in multimodal models. Combining behavioral audits with causal interventions clarifies the problem and yields concrete, testable mitigation strategies, while emphasizing that measured repairability is partial and must be validated across tasks, models, and deployment contexts.

8. Next steps and open questions

We outline a focused roadmap to move from diagnosis to principled repair.

1. Mitigation experiments. Systematically evaluate the following families:

- *Surgical edits.* Head-level fine-tuning, low-rank updates, or soft-pruning of high-leverage attention heads discovered by patching. Measure trade-offs in IAR/SCS and utility metrics.
- *Projection calibration.* Add auxiliary contrastive or margin losses at pooled/projection stages to bias projection geometry toward literal perception–text alignment.
- *Run-time verifiers.* Integrate a lightweight CLIP-style verifier to veto or re-score assertions, and explore abstention thresholds and human-in-the-loop fallbacks.

2. Large-scale replication and generalization. Scale patching across more checkpoints, architectures, and alignment recipes to quantify universality and donor dependence. Report systematic trade-offs (e.g., IAR reduction vs fluency loss).

3. Causal chain-of-effect mapping. Move beyond single-module transplants to multi-stage tracing: after successful transplants, track activation flows through subsequent layers to test whether repairs produce persistent routing changes or

transient nudges. Automated circuit discovery approaches [22] could help systematically map these multi-component pathways.

4. Benchmark and tooling. Release a turn-key MATS benchmark and a reproducible patching toolkit for community evaluation and red-teaming.

These directions will test whether minimally invasive repairs are sufficient for deployment or whether a deeper alignment redesign is necessary.

9. Conclusion

We introduce MATS, a compact behavioral audit, and use large-scale activation patching to localize a concerning multimodal failure mode — pathological truth bias. Behavioral metrics show instruction-tuned generative VLMs frequently affirm visually false statements ($IAR \approx 75\text{--}80\%$) and fail to invert logical predicates ($SCS \approx 1\text{--}3\%$), while contrastive encoders are substantially more robust. Mechanistic probes implicate sparse cross-attention heads in generative models and pooled/projection components in contrastive models as promising repair targets. We release the MATS codebase and encourage the community to adopt audit-first practices and targeted mitigation experiments to make multimodal systems more truthful and verifiable.

References

- [1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1
- [2] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines for visual instruction tuning. In *CVPR*, 2024. 1, 5
- [3] 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. *arXiv preprint arXiv:2308.12966*, 2023. 1
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Ilya Sutskever, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 5
- [5] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language-image pre-training. In *ICCV*, 2023. 1
- [6] Khai-Nguyen Nguyen, et al. Vision language models are biased: A study on object counting and identification. *arXiv preprint arXiv:2505.23941*, 2025. 1
- [7] Jiaqi Wang, Yifei Ming, Zhenfang Shi, Vibhav Vineet, et al. Is a picture worth a thousand words? Delving into spatial reasoning for vision-language models. *arXiv preprint arXiv:2406.14852*, 2024. 1, 2
- [8] Mert Yuksekogonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023. 1, 2, 8
- [9] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022. 2, 3
- [10] Zhiqiang Chen, Yihuai Lan, Guangyi Chen, et al. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 2
- [11] Tristan Thrush, Ryan Jiang, Max Bartolo, et al. Winoground: Probing vision and language models for visio-linguistic compositionality. *arXiv preprint*, 2022. 2
- [12] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, et al. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS*, 2023. 2
- [13] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Findings of EMNLP*, 2023. 2
- [14] Zixian Ma, Jerry Hong, Mustafa Omer Gul, et al. CREPE: Can vision-language foundation models reason compositionally? In *CVPR*, 2023. 2
- [15] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. 1, 2
- [16] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022. 1, 2
- [17] Kevin Wang, Alexandre Variengien, Arthur Conmy, et al. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022. 1, 2
- [18] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Yonatan Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020. 1, 2
- [19] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, 2021. 2
- [20] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, et al. Causal scrubbing: A method for rigorously testing interpretability hypotheses. *arXiv preprint arXiv:2205.10256*, 2022. 2
- [21] Neel Nanda, Lawrence Chan, Tom Lieberum, et al. Progress measures for grokking via mechanistic interpretability. In *ICLR*, 2023. 2
- [22] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, et al. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023. 9
- [23] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 1, 2
- [24] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017. 1, 2, 5, 8
- [25] Nisan Stiennon, Long Ouyang, Jeff Wu, et al. Learning to summarize from human feedback. In *NeurIPS*, 2020. 1, 2, 5, 8

- [26] Manu Sharma, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023. 1, 2, 8
- [27] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022. 1, 2, 8
- [28] Lewei Wang, Wei Chen, Tao Yang, et al. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. In *Transactions of the Association for Computational Linguistics*, 2024. 8
- [29] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023. 2, 8
- [30] Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 2, 5
- [31] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv preprint*, 2023. 2
- [32] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 1, 2
- [33] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, et al. Object hallucination in image captioning. In *EMNLP*, 2018. 1, 2
- [34] Wenzuan Zhou, Sheng Zhang, Hoifung Poon, et al. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 1, 2
- [35] Arjun Gunjal, et al. Mitigating image captioning hallucinations in vision-language models. *arXiv preprint arXiv:2505.03420*, 2025. 1, 2
- [36] Tianyu Huang, et al. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *arXiv preprint arXiv:2504.13169*, 2025. 1, 2
- [37] Marco Favero, et al. Multi-modal hallucination control by visual information grounding. *arXiv preprint arXiv:2403.14003*, 2024. 1, 2
- [38] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *ACL*, 2020. 1, 2
- [39] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Findings of ACL*, 2022. 8