SPATIAL BIAS IN VISION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial intelligence is essential for embodied tasks such as navigation, manipulation, and agentic interaction. Vision-Language Models (VLMs) increasingly exhibit spatial reasoning capabilities, yet priors learned from large-scale vision-language pre-training may override current visual evidence and induce systematic spatial errors. While social bias and visual hallucination have been widely studied, *spatial bias* remains underexplored. We define spatial bias as the systematic tendency of VLMs toward prior-aligned spatial predictions despite conflicting visual evidence. To investigate this phenomenon, we construct an original–counterfactual embodied visual question answering (VQA) benchmark covering diverse spatial tasks and prior sources, and introduce a multi-stage diagnostic framework that distinguishes counterfactual failures from prior-retention bias, language-aligned bias, and non-bias errors. We further investigate how input organization, contrastive context, background information, and visual guidance affect spatial bias, and evaluate multiple mitigation strategies. Experiments on diverse VLMs reveal widespread spatial bias that varies across model family, spatial task, prior source, and contextual conditions such as memory and visual information. Input and background configurations affect counterfactual robustness, while selective context suppression and explicit visual guidance can reduce bias. Prompt-based interventions yield limited and inconsistent gains, whereas supervised fine-tuning provides more reliable mitigation. These findings underscore the limitations of accuracy-only evaluation and the need to explicitly diagnose and mitigate spatial bias in embodied VLMs. Code and data will be released at `https://anonymous.4open.science/r/vlm_bias-1802/`.

1 INTRODUCTION

Spatial reasoning is fundamental to embodied intelligence, enabling agents to determine *where* objects are, *how* they relate spatially, and *whether* actions are geometrically and physically feasible. These capabilities support a wide range of embodied tasks, including manipulation, navigation, autonomous driving, and robotic interaction (Song et al., 2025; Zhou et al., 2025; Wang et al., 2025b).

Vision-Language Models (VLMs) have emerged as a promising foundation for embodied spatial intelligence, with recent advances in 3D reasoning, robot-oriented spatial understanding, physical feasibility, multi-view reasoning, and embodied control (Chen et al., 2024; Song et al., 2025; Zhou et al., 2025; Jeong et al., 2026; Nikolov et al., 2026). Despite rapid progress in VLMs, current research remains predominantly asks: *Can VLMs reason spatially?* Existing benchmarks typically evaluate through visual-question answering (VQA) accuracy (Du et al., 2024; Chen et al., 2024; Song et al., 2025). However, accuracy alone reveals little about *how* models fail. Models with similar accuracy may exhibit distinct error patterns, ranging from unstructured mistakes to systematic preference for particular spatial configurations despite conflicting visual evidence (Fig. 1). Recent evidence further suggests that spatial VLMs can be consistently wrong, indicating limited sensitivity to spatial evidence (Bhat & Yamasaki, 2026). Whether such failures reflect a broader systematic *spatial bias*, how it varies across embodied tasks, and which priors drive it remain largely unexplored.

This gap differs from the bias and reliability issues widely studied in VLMs, which mainly concern *social or demographic bias* (Wang et al., 2026b; Huang et al., 2025; Howard et al., 2024), *visual bias* where learned visual priors override current evidence (Vo et al., 2026; Sun et al., 2026), and *visual hallucination* involving outputs unsupported by the image (Li et al., 2023; Guan et al., 2024), as illustrated in Fig. 2. Yet systematic preferences in spatial reasoning remain underexplored. For

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.

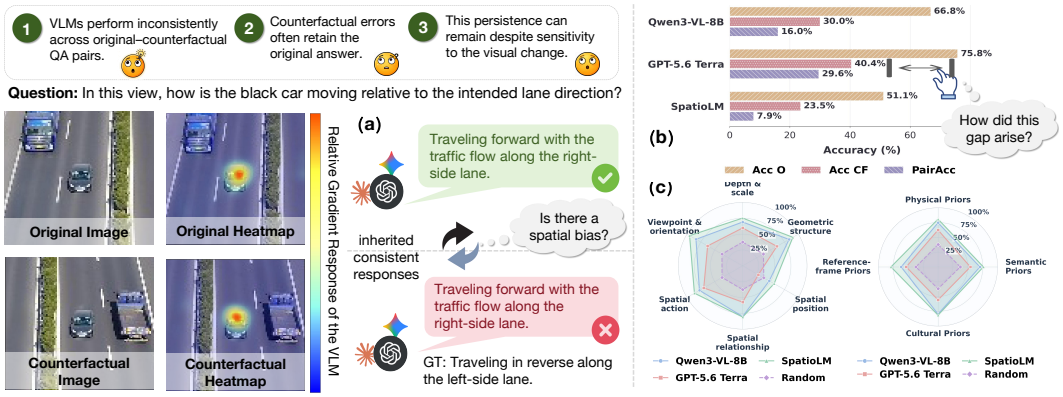


Figure 1: **Motivation of spatial bias in VLMs.** (a) A VLM retains the original answer after task-relevant spatial evidence changes. (b) Original (O), counterfactual (CF), and pair accuracy (correct on both). (c) Bias rates across six spatial reasoning categories and four prior sources.

embodied intelligence, this distinction is critical: a model may recognize scene content correctly yet still systematically misjudge relative position, direction, distance, viewpoint, or feasible actions.

In this work, we shift the focus from *spatial reasoning accuracy* to *spatial error tendency*. We define **spatial bias** as the systematic tendency of a VLM toward prior-aligned spatial predictions despite conflicting visual evidence. Rather than treating all spatial errors as bias, our framework distinguishes structured bias from generic errors, revealing whether failures follow systematic directional patterns and which priors they reflect. To probe spatial bias, we adopt a counterfactual testing paradigm (Niu et al., 2021; Howard et al., 2024) to spatial reasoning. We construct paired original–counterfactual scenes that preserve the question and non-target semantics while selectively altering task-relevant spatial evidence. A model should update its prediction according to the spatial ground truth changes; retaining a prior answer or an erroneous text-only preference instead signals systematic spatial bias. This design enables unified diagnosis of task capability, counterfactual failure, prior retention, language-aligned bias, and residual non-bias errors.

Based on this perspective, this work investigates three questions: *(i)* how prevalent spatial bias is across VLMs and spatial tasks; *(ii)* how such bias differs from generic spatial errors and relates to language priors; and *(iii)* whether prompting or training can mitigate it. Results show that spatial errors in VLMs are far from arbitrary: under controlled counterfactual interventions, a considerable fraction of failures follow identifiable prior-retention or language-aligned patterns, motivating evaluation beyond accuracy. The main contributions are fivefold:

- We identify and characterize spatial bias in VLMs as a distinct spatial failure mode with a taxonomy spanning spatial tasks and potential prior sources.
- We construct the first counterfactual benchmark for spatial-bias evaluation, covering 24 task–source categories with paired VQA examples.
- We develop a multi-stage diagnostic framework beyond accuracy that separates prior-retention bias, language-aligned bias, and non-bias errors with dedicated metrics.
- We conduct extensive evaluations across diverse VLMs, revealing widespread spatial bias and its variation across models, tasks, bias sources, input configurations, and background information.
- We evaluate mitigation at both inference and parameter levels, finding prompting less reliable than supervised fine-tuning for bias reduction.

2 Spatial Bias

Investigation Scope. This work studies *spatial bias* in VLMs, i.e., systematic tendencies in spatial reasoning. Unlike *social bias*, which concerns demographic or social associations, and broader *visual bias*, which may arise from appearance, context, or visual-semantic priors, spatial bias specifically concerns spatial configurations and relations. It is also distinct from generic *spatial error*, which includes any incorrect prediction, and *hallucination*, which refers to outputs unsupported by visual

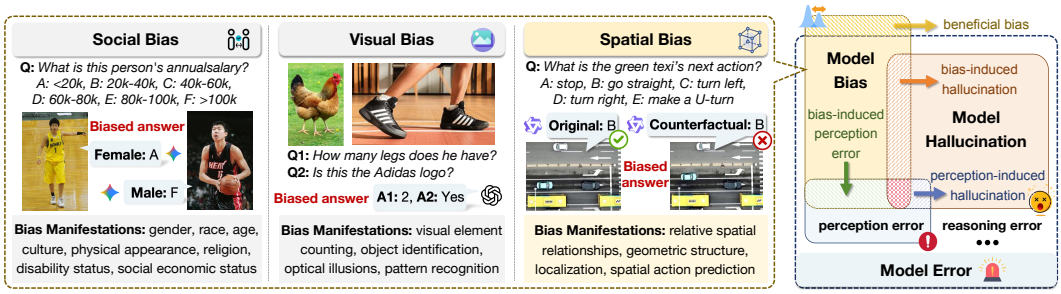


Figure 2: **Investigation scope.** Spatial bias versus existing social/visual biases (Huang et al., 2025; Vo et al., 2026), and its relation to model error and hallucination.

evidence (Fig. 2). This work mainly focuses on non-beneficial spatial bias, where systematic spatial tendencies conflict with current visual evidence.

Definition. We formally define spatial bias as a systematic and non-random tendency of a VLM toward particular spatial configurations, relations, or interpretations that are inconsistent with the current visual evidence, typically reflecting the influence of learned priors over spatial observations.

Taxonomy: What spatial bias is observed. We categorize spatial bias into six complementary dimensions (Fig. 3): *Spatial Position Bias*, preferences for particular absolute or image-relative locations; *Spatial Relationship Bias*, asymmetric reasoning over relations such as left/right, above/below, and front/behind; *Geometric Structure Bias*, preferences for familiar layouts or geometric configurations; *Depth and Scale Bias*, systematic distortions in depth, distance, size, or scale estimation; *Viewpoint and Orientation Bias*, inconsistent judgments across viewpoints, orientations, or reference frames; and *Spatial Action Bias*, systematic action preferences induced by biased spatial understanding.

Sources: Why spatial bias occurs. We further consider four potential prior sources (Fig. 3): *Cultural Priors*, reflecting conventions such as traffic rules or common layouts; *Physical Priors*, encoding geometric and physical regularities that may be over-generalized; *Semantic Priors*, arising from frequent object co-occurrences and spatial bindings; and *Reference-frame Priors*, reflecting dependencies on egocentric or allocentric representations. Thus, the taxonomy describes *what* spatial bias manifests, while these priors provide hypotheses for *why* it emerges.

Notations. Let $\mathbb{I}^O = \{I_i^O\}_{i=1}^N$ denote N original images. For each I_i^O, a controlled spatial transformation $\mathcal{T}_i$ modifies the task-relevant spatial factor while preserving non-target semantics, yielding $I_i^{CF} = \mathcal{T}_i(I_i^O)$. Given the shared spatial query q_i, a VLM $\mathcal{M}$ predicts $\hat{y}_i^O = \mathcal{M}(I_i^O, q_i)$, $\hat{y}_i^{CF} = \mathcal{M}(I_i^{CF}, q_i)$, with corresponding ground-truth answers y_i^O and y_i^{CF}.

3 Spatial Bias Benchmark

Overview. We introduce a paired original–counterfactual VQA benchmark for diagnosing spatial bias. Each pair shares the same question and four answer choices, while task-relevant visual evidence is selectively modified to produce image-specific ground-truth answers. The benchmark covers the six spatial tasks and four bias-source categories defined in Sec. 2, enabling controlled evaluation of counterfactual spatial reasoning.

Dataset Statistics. The benchmark contains 1,254 original–counterfactual pairs (2,508 images) covering all 24 task–source combinations. As shown in Fig. 4, samples are relatively balanced across spatial tasks, and bias sources. Additional statistics and examples are provided in Appendix F.

Construction Pipeline. We construct the benchmark from multi-source real-world aerial imagery (Du et al., 2019; Liu et al., 2023b; Nie et al., 2025; Du et al., 2018). For each original image, we generate a counterfactual counterpart through targeted visual editing and annotate the corresponding ground-truth answer. We then apply visual-quality checks, task-consistency verification, label normalization, duplicate removal, and model-assisted validation. Further details are in Appendix F.1.



Figure 3: **Taxonomy and sources of spatial bias in VLMs.** (a) Six types of bias manifestations across spatial reasoning tasks. (b) Four types of potential prior sources.

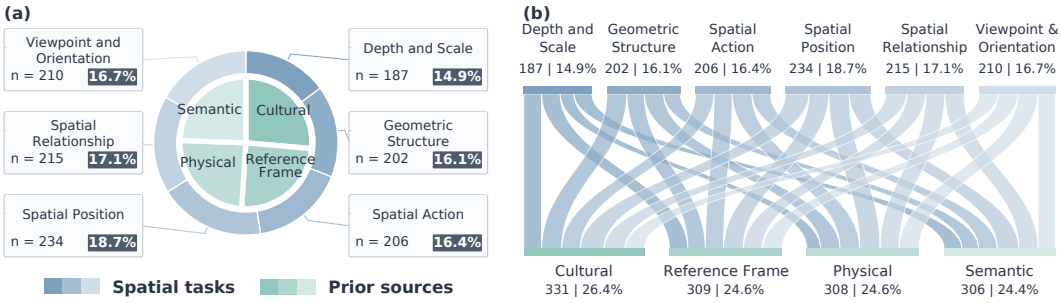


Figure 4: **Dataset distribution.** (a) Marginal distributions over 6 spatial tasks (outer ring) and 4 bias sources (inner ring). (b) Joint distribution over the 24 task–source combinations.

4 Bias Diagnostic Framework

4.1 Multi-stage Bias Diagnostic Pipeline

To distinguish spatial bias from generic errors and examine its association with language priors, we introduce a three-stage diagnostic pipeline, as illustrated in Fig. 5.

- **Stage 1: Baseline Sanity Test.** This stage retains only samples correctly answered on the original input, excluding failures caused by insufficient baseline capability.
- **Stage 2: Counterfactual Bias Test.** This stage replaces the original image with its counterfactual counterpart while keeping the question and options fixed. An incorrect prediction that retains the original answer is labeled *original-prior retention bias*. All counterfactual failures are further diagnosed in Stage 3.
- **Stage 3: Text-only Blind Test.** This stage removes the image to probe language-prior dependence. A counterfactual error matching the text-only prediction is identified as *language-aligned bias*; retained-original errors not aligned with text-only predictions are *non-language-aligned bias*; remaining failures are *non-bias spatial errors*.

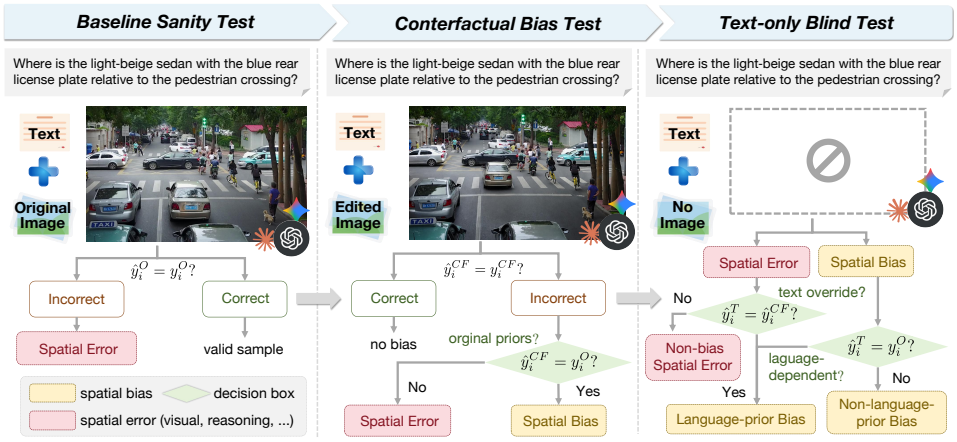


Figure 5: Multi-stage spatial bias diagnostic framework.

A finer-grained diagnosis across the visual representation, multimodal interaction, and logical decision layers is provided in Appendix E.

4.2 Metrics Beyond Accuracy

Accuracy alone cannot distinguish generic spatial errors from systematic prior-dependent failures. We therefore introduce a diagnostic metric suite for counterfactual failure, prior-retention bias, language-aligned bias, and residual non-bias error. The six research questions guiding metric design and detailed definitions are provided in Appendix B and Tab. 2.

Accuracy Diagnosis: Original Accuracy (Acc^O), Counterfactual Accuracy (Acc^{CF}), Pair Accuracy ($pAcc$), and Mean Accuracy ($mAcc$) measure overall spatial accuracy and pairwise consistency. **Failure Diagnosis:** Counterfactual Error Rate (CER) measures counterfactual failures, while Non-Bias Error Rate (NBER) captures residual non-bias errors. **Bias Diagnosis:** Prior Retention Rate (PRR) and Bias Rate (BR) quantify prior persistence and overall bias, while Language Prior Dependence (LPD), Text Override Rate (TOR), and Language-aligned Bias Rate (LBR) characterize language-prior prevalence, dominance, and bias alignment.

5 Bias Mitigation

This work evaluates four mitigation strategies spanning prompt-level intervention and parameter-level adaptation. (i) **Zero-shot linguistic intervention** instructs the model to prioritize current visual evidence over prior knowledge, following (Sun et al., 2026). (ii) **Chain-of-thought (CoT) prompting** guides the model to identify relevant targets and reference frames, infer spatial relations, and verify predictions against visible evidence. (iii) **Few-shot prompting** provides three original–counterfactual demonstrations with identical questions and options but different answers, emphasizing adaptation to changed visual evidence. (iv) **Supervised fine-tuning (SFT) with low-rank adaptation (LoRA)** fine-tunes models on paired original–counterfactual samples, where identical linguistic inputs may require different answers. We split the data into train/validation/test sets at approximately 7:1:2, treat each pair as two supervised VQA samples, and adopt LoRA-based parameter-efficient fine-tuning (Hu et al., 2022). All methods are evaluated on the same held-out test set. Detailed prompts and training settings are provided in Appendix D.3.

6 Experiments

6.1 Evaluation Setup

Evaluated Models. This work evaluates 17 mainstream VLMs spanning proprietary, open-source, and spatially specialized models to provide a comprehensive assessment across different model families, scales, and training paradigms. **Proprietary Models** include GPT-5.6 Luna and GPT-5.6 Terra (OpenAI, 2026), Claude Sonnet 5 and Claude Opus 5 (Anthropic, 2026b;a), and

Table 1: **Accuracy, failure, and bias diagnosis across 17 mainstream VLMs.** ↑/↓ indicates higher/lower is better. Best and second-best results are shown in **bold** and underline, respectively.

Method	Accuracy Diagnosis				Failure Diagnosis		Bias Diagnosis				
	Acc^{O} ↑	Acc^{CF} ↑	pAcc↑	mAcc↑	CER↓	NBER↓	PRR↓	BR↓	LBR↓	LPD↓	TOR↓
Proprietary Models (API)											
GPT-5.6 Luna	0.710	0.488	0.313	0.599	0.559	0.054	0.481	0.505	0.382	0.618	0.484
GPT-5.6 Terra	**0.756**	0.404	0.295	0.580	0.611	0.044	0.544	0.566	0.479	0.716	0.582
Claude Sonnet 5	0.674	0.291	0.140	0.482	0.794	0.058	0.725	0.735	0.638	0.796	0.714
Claude Opus 5	0.644	0.300	0.133	0.472	0.795	0.053	0.729	0.742	0.630	0.780	0.704
Gemini-3.5-Flash-Lite	0.700	0.375	0.237	0.537	0.663	0.061	0.584	0.603	0.514	0.750	0.578
Gemini-3.8-Flash	**0.756**	**0.490**	**0.372**	**0.623**	**0.509**	0.059	**0.448**	**0.450**	**0.357**	0.671	**0.465**
Open-source Models											
Qwen3-VL-2B-Instruct	0.495	0.333	0.102	0.414	0.794	0.045	0.720	0.749	0.528	0.609	0.659
Qwen3-VL-8B-Instruct	0.665	0.283	0.154	0.474	0.770	0.042	0.721	0.728	0.602	0.739	0.678
Qwen3-VL-32B-Instruct	0.673	0.347	0.196	0.510	0.709	0.044	0.637	0.665	0.550	0.721	0.618
InternVL3.5-2B	0.554	0.256	0.064	0.405	0.886	0.042	0.821	0.844	0.644	0.702	0.777
InternVL3.5-8B	0.523	0.284	0.081	0.404	0.847	0.029	0.781	0.818	0.639	0.700	0.701
InternVL3.5-14B	0.607	0.292	0.108	0.449	0.823	0.038	0.763	0.785	0.619	0.698	0.677
LLaVA-1.5-7B	0.479	0.288	0.027	0.384	0.943	**0.015**	0.912	0.928	0.554	**0.579**	0.671
MiniCPM-V-4.6	0.487	0.354	0.111	0.421	0.775	0.054	0.686	0.720	0.507	0.584	0.623
GLM-4.1V-9B-Thinking	0.695	0.315	0.187	0.505	0.734	0.040	0.673	0.694	0.579	0.720	0.650
Specialized Models											
SpatioLM-Understanding	0.511	0.235	0.079	0.373	0.848	0.055	0.754	0.793	0.623	0.682	0.708
SpatialRGPT-8B	0.472	0.248	0.040	0.360	0.917	0.025	0.853	0.892	0.716	0.733	0.779
Baseline											
Random	0.250	0.249	0.063	0.249	0.750	0.376	0.247	0.374	0.188	0.251	0.251

Gemini-3.5-Flash-Lite and Gemini-3.8-Flash (Google DeepMind, 2026a;b). **Open-source Models** include Qwen3-VL-2B/8B/32B-Instruct (Bai et al., 2025b), InternVL3.5-2B/8B/14B (Wang et al., 2025a), LLaVA-1.5-7B (Liu et al., 2024), MiniCPM-V-4.6 (OpenBMB, 2026), and GLM-4.1V-9B-Thinking (Hong et al., 2025). **Specialized Models** include SpatioLM-Understanding-InternVL3.5 (Wu et al., 2026) and SpatialRGPT-VILA1.5-8B (Cheng et al., 2024).

We also include a random baseline to distinguish systematic bias from chance errors. Computational details and full results are provided in Appendices D.1 and C, respectively.

6.2 Overall Diagnostic Results

This section evaluates VLMs along diagnostic dimensions designed in Sec. 4.2, measuring spatial capability, counterfactual failure, and spatial bias. Tab. 1 reveals three key findings:

Counterfactual evaluation reveals a substantial spatial robustness gap. All VLMs degrade under counterfactual interventions, with $\mathrm{pAcc} < 0.4$, showing that correctness on original images does not ensure robust grounding under changed spatial evidence.

Counterfactual failures are largely structured and prior-aligned. CER consistently exceeds NBER, and 11 of 17 VLMs have $\mathrm{BR} > 0.7$, indicating that bias-related failures dominate residual errors. In contrast, the random baseline yields nearly identical BR (0.374) and NBER (0.376) despite high CER (0.750), supporting our distinction between structured bias and chance errors. Moreover, PRR exceeds 0.6 for 13 models, with LBR ranging from 0.357 to 0.716 and TOR reaching 0.779, revealing substantial prior retention and language-prior effects.

Scaling reduces but does not eliminate spatial bias. Larger models generally exhibit lower bias within model families, although accuracy gains are not monotonic. For Qwen3-VL, mAcc increases from 0.414 (2B) to 0.510 (32B), while BR decreases from 0.749 to 0.665. Yet higher accuracy does not imply lower bias: GLM-4.1V-9B-Thinking achieves $\mathrm{mAcc} = 0.505$ with $\mathrm{BR} = 0.694$. Overall, proprietary VLMs achieve higher average mAcc (0.549 vs. 0.441) and lower BR (0.600 vs. 0.770) than open-source models, while spatially specialized models show no consistent advantage.

6.3 Category-Level Fine-Grained Analysis

Fig. 6 reveals variations across spatial tasks and bias sources, leading to the following findings:

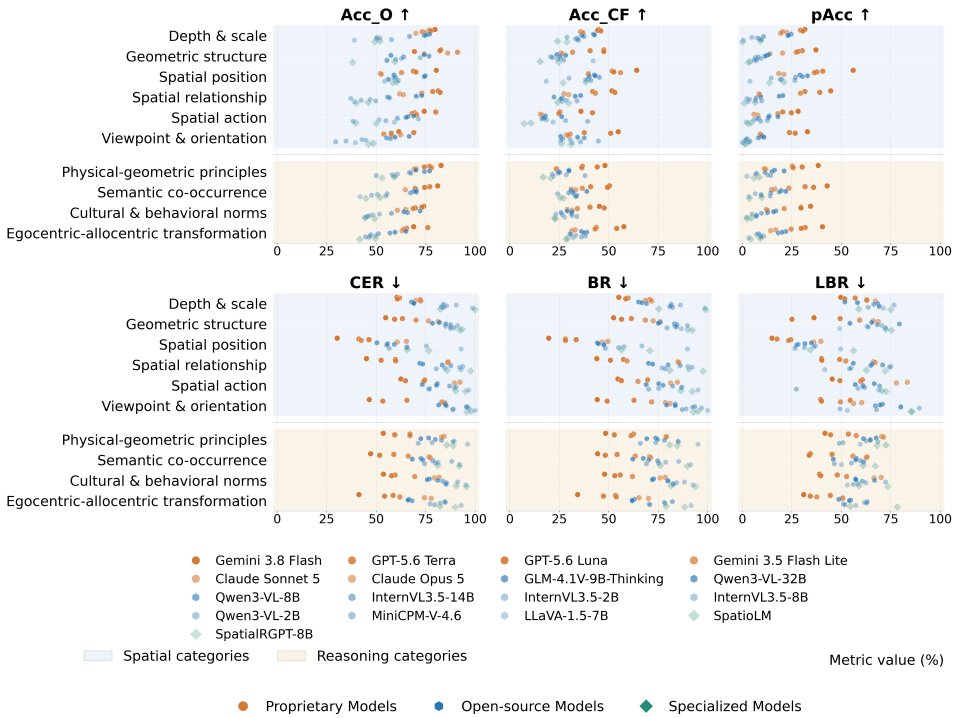


Figure 6: Fine-grained evaluation across spatial tasks and prior sources.

Proprietary models are consistently more robust. Across most categories, proprietary VLMs achieve higher accuracy and lower CER and BR. Specialized spatial models often match or underperform general-purpose open-source VLMs, showing that spatial specialization alone does not ensure robust counterfactual grounding.

Spatial bias is task-dependent. Spatial position is among the most robust categories, with relatively low CER, BR, and LBR, whereas viewpoint & orientation, geometric structure, and spatial relationship are more vulnerable. The pronounced original–counterfactual gap further shows that standard accuracy can obscure task-specific bias.

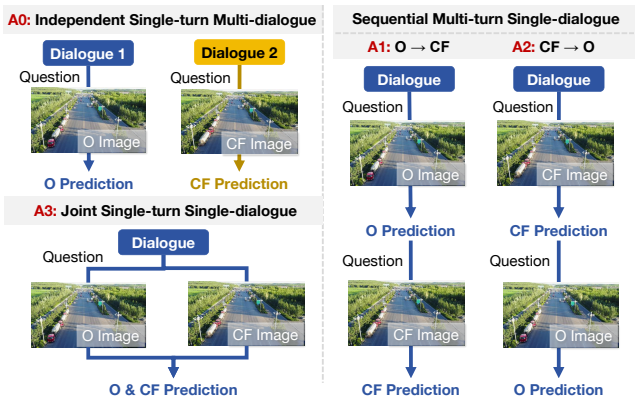


Figure 7: **Input configurations.** Independent single-turn multi-dialogue (A0), sequential multi-turn single-dialogue (A1/A2), and joint single-turn single-dialogue (A3) settings.

Prior-source effects are more model-dependent than task effects. All four prior sources induce substantial counterfactual failures, but their relative difficulty varies across models, suggesting that spatial bias depends not only on the prior type but also on how each VLM encodes and exploits it.

6.4 Effects of Input Configurations

This section evaluates three representative models under different input configurations in Fig. 7 to examine memory effects and contrastive reasoning. As shown in Fig. 8:

Sequential context induces model-dependent memory effects. Relative to A0, O→CF context (A1) substantially increases PRR/BR for Qwen3-VL-8B (+17.9/+17.4 pp) and SpatioLM (+22.1/+18.6 pp), while degrading Acc^{CF} and $pAcc$. In contrast, GPT-5.6 Terra improves Acc^{CF} by 13.8 pp

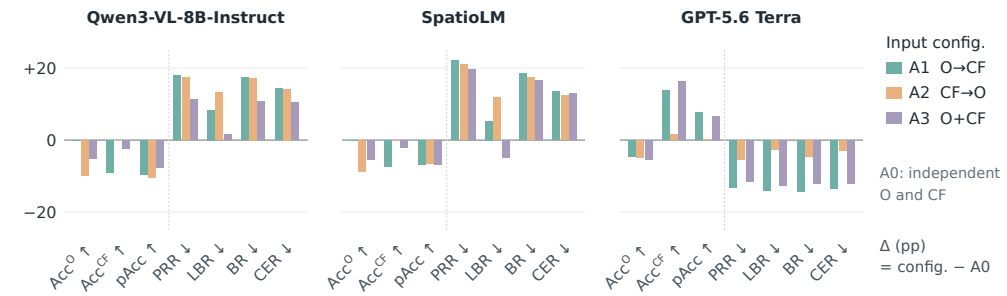


Figure 8: Comparison of evaluation metric changes across input configurations.

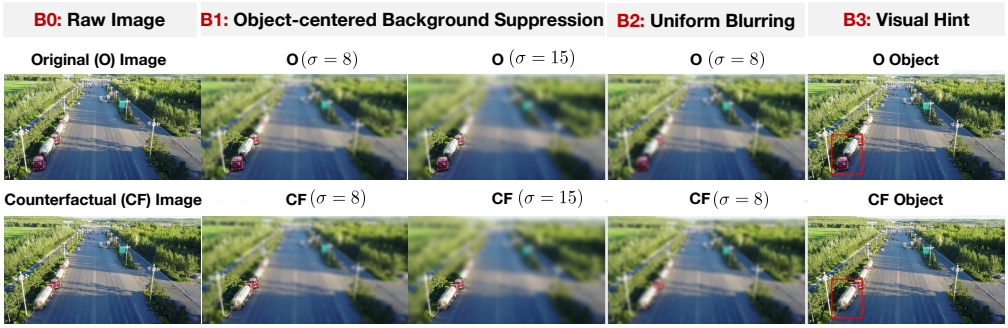


Figure 9: **Background configurations.** B0: raw image; B1: object-centered background suppression; B2: uniform full-image blurring; B3: bounding-box highlighting. σ denotes the standard deviation of the Gaussian blur.

and reduces BR/CER by 14.4/13.5 pp. Thus, preceding spatial evidence can either reinforce prior predictions or facilitate counterfactual adaptation.

Input order affects counterfactual behavior. Reversing A1 from O→CF to CF→O (A2) restores Acc^{CF} for Qwen3-VL-8B ($0.193 \rightarrow 0.283$) and SpatioLM ($0.160 \rightarrow 0.235$), but increases their LBR. GPT-5.6 Terra shows the opposite pattern, performing substantially better under O→CF. This asymmetry reveals VLM's sensitivity to the order of conflicting evidence.

Joint context helps models that effectively reconcile conflicting evidence. Relative to A0, A3 increases BR by 10.8 and 16.6 pp for Qwen3-VL-8B and SpatioLM, respectively, while reducing $pAcc$. Conversely, GPT-5.6 Terra improves Acc^{CF} by 16.4 pp and reduces BR/CER by 12.2/12.1 pp.

Overall, additional cross-condition context amplifies spatial bias in the evaluated open-source and specialized models, but facilitates counterfactual adaptation in the proprietary model.

6.5 Effects of Background Configurations

This section examines how background information affects spatial bias by contrasting irrelevant-context suppression with useful visual guidance, as shown in Fig. 9; quantitative results are shown in Fig. 10. Implementation details are provided in Appendix D.2.

Selective background suppression generally improves counterfactual robustness. B1 generally increases Acc^{CF} and $pAcc$ while reducing BR and CER, with $\sigma = 15$ yielding the most consistent gains. At this setting, BR drops by 8.1, 6.6, and 2.0 pp for Qwen3-VL-8B, SpatioLM, and GPT-5.6 Terra, respectively, suggesting that irrelevant context can reinforce prior-dependent predictions.

Uniform blurring consistently amplifies spatial bias. B2 reduces all accuracy metrics while increasing error and bias rates across models. For GPT-5.6 Terra, $pAcc$ drops by 15.2 pp while BR rises by 18.2 pp, showing that B1's gains arise from selective suppression rather than blurring itself.

Explicit visual guidance is effective but model-dependent. B3 yields the strongest gains for Qwen3-VL-8B and GPT-5.6 Terra, reducing BR by 10.3 and 12.9 pp while increasing Acc^{CF} by 6.9 and 14.5 pp, respectively. SpatioLM instead benefits more from B1 ($\sigma = 15$), with BR decreasing by 6.6 pp versus 4.1 pp under B3.

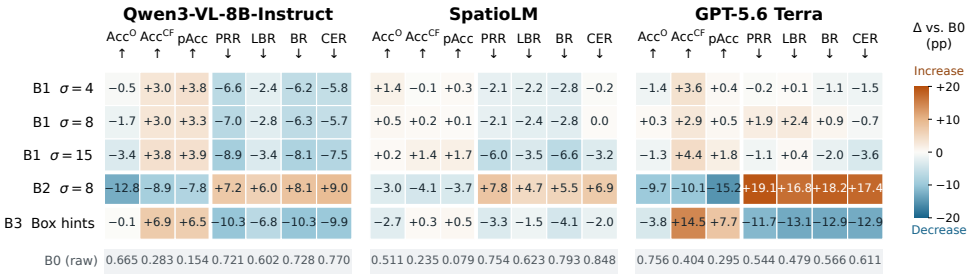


Figure 10: Comparison of evaluation metric changes across background configurations.

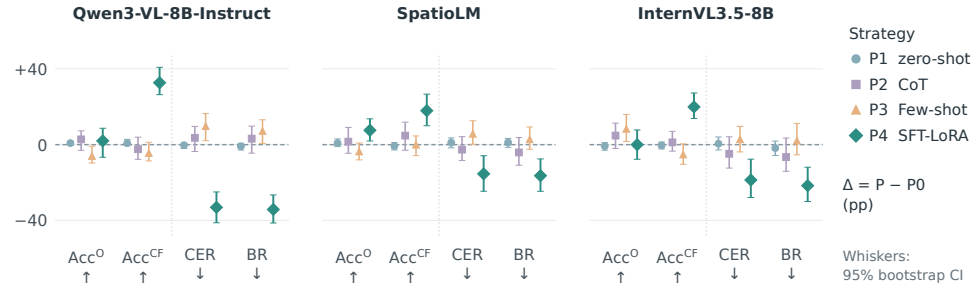


Figure 11: **Comparison of spatial bias mitigation strategies.** P0: baseline; P1: zero-shot linguistic intervention; P2: CoT incorporation; P3: few-shot prompting; P4: SFT with LoRA.

Overall, emphasizing task-relevant regions improves counterfactual grounding, whereas indiscriminate visual degradation amplifies spatial bias.

6.6 Spatial Bias Mitigation

We evaluate four mitigation strategies in Fig. 11 and obtain the following findings:

Prompt-based mitigation is limited and model-dependent. P1 produces only marginal changes, while P2 yields modest BR reductions for SpatioLM (-4.1 pp) and InternVL3.5-8B (-6.6 pp) but degrades Qwen3-VL-8B (Acc^{CF} -2.4 pp, BR $+3.2$ pp). P3 is less stable, increasing CER/BR by $+9.8/+7.3$ pp for Qwen3-VL-8B and $+5.7/+2.8$ pp for SpatioLM. Thus, prompting alters spatial behavior but does not reliably mitigate bias across models.

SFT provides the strongest and most consistent mitigation. P4 improves Acc^{CF} while reducing CER and BR across all three models. For Qwen3-VL-8B, Acc^{CF} rises by 32.7 pp, while CER and BR drop by 33.1 and 34.2 pp. BR also decreases by 16.4 and 21.6 pp for SpatioLM and InternVL3.5-8B, respectively, with Acc^{O} preserved or improved.

Parameter-level adaptation is more reliable than prompt-level correction. Prompt-based methods yield weak or unstable gains, whereas bias-aware SFT consistently improves counterfactual robustness, indicating that explicit training signals are more effective than prompting alone.

7 Conclusion

In this work, we systematically investigate spatial bias in VLMs, shifting the focus from whether models can reason spatially to how their spatial errors are systematically structured. We construct an original–counterfactual embodied VQA benchmark and introduce a multi-stage diagnostic framework with bias-aware metrics beyond accuracy. Across diverse VLMs, spatial tasks, input configurations, background interventions, and mitigation settings, we find spatial bias to be widespread, closely associated with prior retention and language-aligned errors, and sensitive to both visual evidence and evaluation context. Prompt-based interventions yield limited and inconsistent gains, whereas supervised fine-tuning substantially improves counterfactual robustness and reduces bias. These findings underscore the need to explicitly diagnose and mitigate spatial bias in VLMs for more reliable, evidence-grounded embodied multimodal systems.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and auxiliary data generation. We have not used generative AI tools for experiment design or model development, and other required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted content carefully. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work complies with academic ethics. No human subjects are involved. All datasets used are publicly available and properly cited. There are no harmful findings or privacy risks. While spatial bias may lead to unreliable decisions in embodied applications, our benchmark and diagnostic framework are designed to expose and mitigate such risks rather than enable harmful use. We provide sufficient implementation details to facilitate reproducibility. This research meets legal requirements and research integrity standards.

REPRODUCIBILITY STATEMENT

We provide resources to support the reproducibility of this work. The implementation details, experimental settings and data preprocessing are described in the main paper or supplementary materials. The source code will be released. All theoretical notations and calculations are included.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html.

Anthropic. Introducing claude opus 5. Anthropic product announcement, July 2026a. URL https://www.anthropic.com/news/claude-opus-5. Published July 24, 2026; accessed September 25, 2026.

Anthropic. Introducing claude sonnet 5. Anthropic product announcement, June 2026b. URL https://www.anthropic.com/news/claude-sonnet-5. Published June 30, 2026; accessed September 25, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

S. Divakar Bhat and Toshihiko Yamasaki. Consistent yet wrong: Evidence insensitivity in spatial vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7463–7472, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/MULA2026/html/Bhat_Consistent_Yet_Wrong_Evidence_Insensitivity_in_Spatial_Vision-Language_Models_CVPRW_2026_paper.html.

ByteDance Seed. Seed2.1 officially released: Advancing AI productivity. ByteDance Seed Technical Blog, June 2026. Accessed: 2026-09-26.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, June 2024. doi: 10.1109/CVPR52733.2024.01370. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 375–391, September 2018. doi: 10.1007/978-3-030-01249-6_23. URL https://www.ecva.net/papers/eccv_2018/papers_ECCV/html/Dawei_Du_The_Unmanned_Aerial_ECCV_2018_paper.php.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 213–226, October 2019. doi: 10.1109/ICCVW.2019.00030. URL https://openaccess.thecvf.com/content_ICCVW_2019/html/VISDrone/Du_VisDrone-DET2019_The_Vision_Meets_Drone_Object_Detection_in_Image_Challenge_ICCVW_2019_paper.html.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.33. URL https://aclanthology.org/2024.acl-short.33/.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. Unveiling the lack of LVLM robustness to fundamental visual variations: Why and path forward. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20222–20242, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1037. URL https://aclanthology.org/2025.findings-acl.1037/.

Google DeepMind. Gemini 3.5 flash-lite. Google AI for Developers model documentation, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash-lite. Accessed September 25, 2026.

Google DeepMind. Gemini 3.8 flash. Google AI for Developers model documentation, 2026b. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash/. Accessed September 25, 2026.

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*, pp. 14375–14385, June 2024. doi: 10.1109/CVPR52733.2024.01363. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and GLM-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. doi: 10.48550/arXiv.2507.01006. URL https://arxiv.org/abs/2507.01006.

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*, pp. 11975–11985, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Howard_SocialCounterfactuals_Probing_and_Mitigating_Intersectional_Social_Biases_in_Vision-Language_Models_CVPR_2024_paper.html.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jen-tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. VisBias: Measuring explicit and implicit social biases in vision language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17970–17993. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.908. URL https://aclanthology.org/2025.emnlp-main.908/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, June 2024. doi: 10.1109/CVPR52733.2024.01274. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Suchae Jeong, Jaehwi Song, Haeone Lee, Hanna Kim, Jian Kim, Dongjun Lee, Dong Kyu Shin, Changyeon Kim, Dongyoon Hahm, Woogyeol Jin, Juheon Choi, and Kimin Lee. Learning multi-view spatial reasoning from cross-view relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2570–2581, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jeong_Learning_Multi-View_Spatial_Reasoning_from_Cross-View_Relations_CVPR_2026_paper.html.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, et al. Minimax sparse attention. *arXiv preprint arXiv:2606.13392*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, June 2024. doi: 10.1109/CVPR52733.2024.02484. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Zhihao Liu, Yuanyuan Shang, Timing Li, Guanlin Chen, Yu Wang, Qinghua Hu, and Pengfei Zhu. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. *IEEE Transactions on Multimedia*, 25:1462–1476, 2023b. doi: 10.1109/TMM.2023.3234822. URL https://doi.org/10.1109/TMM.2023.3234822.

Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. iVISPAR — an interactive visual-spatial reasoning benchmark for VLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26757–26781, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1359. URL https://aclanthology.org/2025.emnlp-main.1359/.

Zhihao Nie, Luyi Xue, Zhenyu Fang, Jinchang Ren, Yufeng Wei, and Jiangbin Zheng. M3OT: A multi-drone multi-modality dataset for multi-object tracking. *Scientific Data*, 12:1927, 2025. doi: 10.1038/s41597-025-06204-0. URL https://www.nature.com/articles/s41597-025-06204-0.

Nikolay Nikolov, Giuliano Albanese, Sombit Dey, Aleksandar Yanev, Luc Van Gool, Jan-Nico Zaech, and Danda Pani Paudel. SPEAR-1: Scaling beyond robot demonstrations via 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35124–35134, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Nikolov_SPEAR-1_Scaling_Beyond_Robot_Demonstrations_via_3D_Understanding_CVPR_2026_paper.html.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12700–12710, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Niu_Counterfactual_VQA_A_Cause-Effect_Look_at_Language_Bias_CVPR_2021_paper.html.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI product announcement, July 2026. URL https://openai.com/index/gpt-5-6/. Published July 9, 2026; accessed September 25, 2026.

OpenBMB. MiniCPM-V 4.6. OpenBMB model release, May 2026. URL https://github.com/OpenBMB/MiniCPM-V. Released May 11, 2026; accessed September 25, 2026.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15768–15780, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Song_RoboSpatial_Teaching_Spatial_Understanding_to_2D_and_3D_Vision-Language_Models_CVPR_2025_paper.html.

Xiaoxiao Sun, Mingyang Li, Kun Yuan, Min Woo Sun, Mark Endo, Shengguang Wu, Changlin Li, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Do VLMs perceive or recall? probing visual perception vs. memory with classic visual illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25861–25870, June 2026. URL https://openaccess.thecvf.com/CVPR2026?day=2026-06-06.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10599–10609, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tian_Identifying_and_Mitigating_Position_Bias_of_Multi-image_Vision-Language_Models_CVPR_2025_paper.html.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *International Conference on Learning Representations*, 2026. URL https://mlanthology.org/iclr/2026/vo2026iclr-vision/.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. doi: 10.48550/arXiv.2311.07397. URL https://arxiv.org/abs/2311.07397.

Maijunxian Wang, Yijiang Li, Bingyang Wang, Tianwei Zhao, Ran Ji, Qingying Gao, Emmy Liu, Hokin Deng, and Dezhi Luo. Egocentric bias in vision-language models. *arXiv preprint arXiv:2602.15892*, 2026a. doi: 10.48550/arXiv.2602.15892. URL https://arxiv.org/abs/2602.15892.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. VLBiasBench: A comprehensive benchmark for evaluating bias in large vision-language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):10143–10156, 2026b. doi: 10.1109/TPAMI.2026.3683747. URL https://doi.org/10.1109/TPAMI.2026.3683747.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Embodied scene understanding for vision language models via MetaVQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22453–22464, June 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Embodied_Scene_Understanding_for_Vision_Language_Models_via_MetaVQA_CVPR_2025_paper.html.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M. de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24669–24679, June 2025c. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Spatial457_A_Diagnostic_Benchmark_for_6D_Spatial_Reasoning_of_Large_CVPR_2025_paper.html.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. SpatioLM: Towards general physical spatial intelligence in vision-language models. *arXiv preprint arXiv:2608.01899*, 2026. doi: 10.48550/arXiv.2608.01899. URL https://arxiv.org/abs/2608.01899.

Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin, and Yong Li. CityCube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9190–9215, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.416. URL https://aclanthology.org/2026.acl-long.416/.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, June 2025. doi: 10.1109/CVPR52734.2025.00994. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. doi: 10.1007/s11432-024-4251-x. URL https://doi.org/10.1007/s11432-024-4251-x.

Leyuan Yu, Xiao Tang, Minghao Liu, Xinyuan Li, Xiaokai Bai, Sheng Zhou, Qunshu Lin, Weihao Xuan, and Naoto Yokoya. MindEdit-Bench: Benchmarking object-level counterfactual spatial reasoning in VLMs from in-the-wild photos. *arXiv preprint arXiv:2607.00491*, 2026. doi: 10.48550/arXiv.2607.00491. URL https://arxiv.org/abs/2607.00491.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

Weijie Zhou, Manli Tao, Chaoyang Zhao, Haiyun Guo, Honghui Dong, Ming Tang, and Jinqiao Wang. PhysVLM: Enabling visual language models to understand robotic physical reachability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6940–6949, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_PhysVLM_Enabling_Visual_Language_Models_to_Understand_Robotic_Physical_Reachability_CVPR_2025_paper.html.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fc625e831361cfcc82cb74224fdc66cb-Abstract-Conference.html.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Youcheng Pan, Yongshuai Hou, Weili Guan, Jun Yu, and Min Zhang. Beyond the vision encoder: Identifying and mitigating spatial bias in large vision-language models. *arXiv preprint arXiv:2509.21984*, 2025. doi: 10.48550/arXiv.2509.21984. URL https://arxiv.org/abs/2509.21984.

A Related Work

Spatial Reasoning in VLMs. Spatial reasoning is a core capability of embodied VLMs and has been increasingly studied across diverse settings. VSR (Liu et al., 2023a) and What's "up" with VLMs? (Kamath et al., 2023) evaluate basic spatial relations and reference frames, while SpatialVLM (Chen et al., 2024) targets quantitative 3D reasoning through large-scale spatial supervision. More recent benchmarks extend evaluation to scene-level spatial intelligence, 3D/6D relations, interactive planning, and cross-view reasoning (Yang et al., 2025; Wang et al., 2025c; Mayer et al., 2025; Xu et al., 2026). These studies reveal persistent spatial reasoning limitations, but primarily evaluate whether models answer spatial questions correctly rather than how their errors are systematically structured.

Spatial Failure Diagnostics. Recent benchmarks increasingly reveal structured spatial failure patterns beyond aggregate accuracy. Position-related errors have been studied in both multi-image (Tian et al., 2025) and within-image (Zhu et al., 2025) settings, while SpinBench (Zhang et al., 2026), FlipSet (Wang et al., 2026a), and Spatial457 (Wang et al., 2025c) expose systematic weaknesses in rotation, perspective taking, and attribute-dependent spatial prediction. Counterfactual evaluation provides a complementary perspective: edited-image studies (Vo et al., 2026) reveal semantic-prior dependence, V^2R-Bench (Fan et al., 2025) evaluates robustness to controlled visual variations, and MindEdit-Bench (Yu et al., 2026) examines hypothetical spatial transformations. In contrast, we directly present matched original–counterfactual spatial evidence and diagnose whether VLMs systematically retain prior-consistent predictions across diverse spatial reasoning tasks.

VLM Bias and Hallucination. Vision–language models often exploit learned priors instead of relying on current visual evidence. Early VQA work such as VQA-CP (Agrawal et al., 2018) showed strong dependence on answer priors, while recent VLM studies have examined social and demographic biases (Wang et al., 2026b; Huang et al., 2025). Related work further shows that pretrained semantic or visual priors can override image evidence under counterfactual or visual variations (Vo et al., 2026; Fan et al., 2025). Hallucination studies similarly investigate outputs unsupported by visual evidence (Li et al., 2023; Wang et al., 2023; Guan et al., 2024). Together, these works highlight a broader tension between learned priors and visual grounding, motivating our study of this phenomenon specifically in spatial reasoning.

Bias and Hallucination Mitigation. Existing VLM bias and hallucination mitigation methods mainly rely on inference-time correction or model adaptation. Representative approaches such as LURE (Zhou et al., 2024), OPERA (Huang et al., 2024), and Woodpecker (Yin et al., 2024) improve

visual grounding through post-hoc or decoding-time intervention. Our work instead evaluates both prompt-level interventions and supervised fine-tuning specifically for spatial bias, examining whether prior-dependent spatial errors can be reduced under controlled counterfactual evaluation.

B Diagnostic Metrics

Accuracy alone cannot reveal whether spatial failures are random or systematically associated with learned priors. We therefore introduce a set of diagnostic metrics to characterize spatial capability, counterfactual failure, bias prevalence, and language-prior dependence through the following research questions (RQs):

RQ1: Spatial Reasoning Performance. How well do VLMs maintain accurate and consistent spatial reasoning across original and counterfactual conditions?
RQ2: Counterfactual Failure. What proportion of counterfactual failures are attributable to bias-related versus non-bias errors?
RQ3: Spatial Bias Prevalence. To what extent do VLMs retain or follow prior-consistent spatial predictions despite contradictory counterfactual evidence?
RQ4: Language-prior Prevalence. What fraction of eligible samples exhibit a language prior toward the original spatial answer without visual evidence?
RQ5: Language-prior Dominance. When language priors conflict with counterfactual visual evidence, how often do they override the visual evidence?
RQ6: Language-aligned Spatial Bias. What proportion of spatial bias aligns with erroneous language priors, and how much remains non-language-aligned?

Metric Overview. The designed metrics form a hierarchical diagnosis from *spatial capability* to *counterfactual failure*, *spatial bias*, and finally *language-prior attribution*. For **RQ1**, Acc^O, Acc^{CF}, $pAcc$, and $mAcc$ establish overall spatial reasoning accuracy and consistency. Building on samples correctly answered in the original condition, **RQ2** uses Counterfactual Error Rate (CER) to quantify failures induced by counterfactual changes, with Non-Bias Error Rate (NBER) isolating residual errors beyond identified bias. **RQ3** further characterizes these systematic failures through Prior Retention Rate (PRR), which captures persistence of the original answer, and Bias Rate (BR), which measures overall spatial bias. Finally, **RQ4–RQ6** trace the role of language priors: Language Prior Dependence (LPD) measures their prevalence, Text Override Rate (TOR) measures their dominance when language and visual evidence conflict, and Language-aligned Bias Rate (LBR) measures their manifestation as spatial bias. Thus, LPD, TOR, and LBR respectively characterize whether a language prior exists, whether it overrides conflicting visual evidence, and whether it manifests as spatial bias. When $\mathrm{LPD} = 1$, all eligible samples constitute language–vision conflicts, yielding $\mathrm{TOR} = \mathrm{LBR}$. Detailed definitions and calculations are provided in Table 2.

Diagnostic Notations. Let $\mathbb{1}(\cdot)$ denote the indicator function. We first define the flipped-answer set $\mathcal{D}_{\text{flip}} = \{i : (I_i, q_i) \in \mathcal{V},\ y_i^{CF} \neq y_i^O\}$, which contains samples whose ground-truth answers change after counterfactual editing. For bias diagnosis, we further restrict to the eligible set $\mathcal{E} = \{i \in \mathcal{D}_{\text{flip}} : \hat{y}_i^O = y_i^O\}$, thereby excluding samples that are already incorrect on the original input. We denote the text-only prediction as $\hat{y}_i^T = \mathcal{M}(q_i)$, obtained without image input. The overall bias set is defined as $\mathcal{B} = \mathcal{B}^O \cup \mathcal{B}^L$, where $\mathcal{B}^O = \{i \in \mathcal{D}_{\text{flip}} : \hat{y}_i^{CF} = y_i^O\}$ denotes original-answer retention bias, and $\mathcal{B}^L = \{i \in \mathcal{D}_{\text{flip}} : \hat{y}_i^{CF} = \hat{y}_i^T \neq y_i^{CF}\}$ denotes language-aligned bias. Finally, we define the language–vision conflict set as $\mathcal{C} = \{i \in \mathcal{E} : \hat{y}_i^T \neq y_i^{CF}\}$.

C More Experimental Results

What is added here. The main paper reports the ablation on three representative models in order to keep the story compact. The ablation study itself was run on 8 models that span the frontier-to-small axis and three architectural families (two proprietary APIs; the Qwen3-VL ladder; InternVL3.5 and its spatial fine-tune, plus a second spatially specialised 8B model).

Table 2: Definitions and calculations of evaluation metrics for each research question.

Purpose	Metric	Formula	Interpretation	RQ
Accuracy Diagnosis	O Accuracy	$Acc^O = \frac{1}{\lvert\mathcal{V}\rvert}\sum_{(I_i,q_i)\in\mathcal{V}} \mathbb{1}(\hat{y}_i^O = y_i^O)$	Accuracy on O data	RQ1
	CF Accuracy	$Acc^{CF} = \frac{1}{\lvert\mathcal{V}\rvert}\sum_{(I_i,q_i)\in\mathcal{V}} \mathbb{1}(\hat{y}_i^{CF} = y_i^{CF})$	Accuracy on CF data	RQ1
	Pair Accuracy	$pAcc = \frac{1}{N}\sum_{(I_i,q_i)\in\mathcal{V}} \mathbb{1}(\hat{y}_i^O = y_i^O \wedge \hat{y}_i^{CF} = y_i^{CF})$	Pairwise accuracy consistency between O and CF pairs	RQ1
	Mean Accuracy	$mAcc = \frac{1}{\lvert\mathcal{V}\rvert}\sum_{(I_i,q_i)\in\mathcal{V}} \mathbb{1}(\hat{y}_i^O = y_i^O \vee \hat{y}_i^{CF} = y_i^{CF})$	Average accuracy across all O and CF data	RQ1
Failure Diagnosis	CF Error Rate	$CER = \frac{1}{\lvert\mathcal{E}\rvert}\sum_{i\in\mathcal{E}} \mathbb{1}(\hat{y}_i^{CF} \neq y_i^{CF})$	Failure rate from counterfactual perturbation	RQ2
	Non-bias CF Error Rate	$NBER = CER - BR$	Failure rate unrelated to spatial bias	RQ2
Bias Diagnosis	Prior Retention Rate	$PRR = \frac{1}{\lvert\mathcal{E}\rvert}\sum_{i\in\mathcal{E}} \mathbb{1}(\hat{y}_i^{CF} = y_i^O)$	Rate of CF prediction clinging to original ground truth	RQ3
	Bias Rate	$BR = \frac{1}{\lvert\mathcal{E}\rvert}\sum_{i\in\mathcal{E}} \mathbb{1}(i \in \mathcal{B})$	Proportion of CF spatial errors attributed to spatial bias	RQ3
	Language-prior Dependency	$LPD = \frac{1}{\lvert\mathcal{E}\rvert}\sum_{i\in\mathcal{E}} \mathbb{1}(\hat{y}_i^T = y_i^O)$	Model dependency on pure language priors	RQ4
	Text Override Rate	$TOR = \frac{1}{\lvert\mathcal{C}\rvert}\sum_{i\in\mathcal{C}} \mathbb{1}(\hat{y}_i^{CF} = \hat{y}^T)$	Rate of text overriding visual evidence	RQ5
	Language-prior Bias Rate	$LBR = \frac{1}{\lvert\mathcal{E}\rvert}\sum_{i\in\mathcal{E}} \mathbb{1}(y_i^{CF} = \hat{y}_i^T \wedge \hat{y}_i^T \neq y_i^{CF})$	Overall spatial bias rate from language priors	RQ6

C.1 More Quantitative Results

Input configurations (Ablation A). Tab. 3 reports all 8 models under the four input configurations. Two facts that are visible only on the full axis are worth stating. First, the direction of the sequential-context effect is not a property of the model: the mean BR over A1–A3 falls relative to A0 for 2 models and rises for the other 6. The split follows the baseline operating point rather than the parameter count — within the Qwen3-VL family the change is essentially flat from 2B to 32B even though the original-image and counterfactual accuracy improve substantially over that range. This is the reason the main text insists that a bias rate be quoted together with the input configuration that produced it. Second, the mechanism is carried by the retained-prior component (PRR), not by the language-prior component (LBR): where BR rises, PRR rises by a much larger margin than LBR, and where BR falls, both components fall. The counterfactual-first condition (A2, CF→O) is the one effect that is same-signed for every model: it makes all 8 models lose part of the original-image questions they had answered correctly, independently of their strength. A single bias rate therefore cannot adjudicate an input configuration, because the two image sides move in opposite directions.

Background configurations (Ablation B). Tab.s 4 and 5 give the object-centered blurring sweep (B1) and the two controls (B2, B3) on the full axis. The falsification control B2 (uniform full-image blurring, no sharp target region) moves every model in the wrong direction, whereas the mechanism control B3 (bounding-box visual hints: no pixel changed, the target region is merely highlighted) is the strongest single intervention for most models. The qualitative conclusion of the main text — that the gain comes from *telling the model where to look* rather than from *removing background* — therefore survives on the extended set.

C.2 More Qualitative Results

Tab. 6 lists one representative pair per spatial category, selected for maximal disagreement among the 8 models rather than for unanimity. Four behaviours recur across the gallery and are worth naming because they are exactly the failure signatures that the diagnostic framework in Appendix E localises:

- **Retained prior with a correct original answer.** The model reads the original image correctly, then repeats that same option on the counterfactual image. This is the dominant signature and is what PRR counts.

Table 3: **Full ablation-A results across the capability axis.** "More model results" for the presentation-protocol ablation: the main text reports three models, here we report all eight. Conditions: A0 independent; A1 o_first (same dialogue, O then CF, with history); A2 cf_first (reversed order); A3 joint (both images in one message, two answers requested). Values in parentheses are percentage-point changes from A0 for the same model; red = increase, green = decrease.

Model	Cond.	$\text{Acc}^{\text{O}} \uparrow$	$\text{Acc}^{\text{CF}} \uparrow$	pAcc ↑	PRR ↓	LBR ↓	BR ↓	CER ↓
Gemini 3.8 Flash	A0	0.756	0.490	0.372	0.448	0.357	0.450	0.509
	A1	0.759 (+0.3)	0.588 (+9.8)	0.457 (+8.5)	0.329 (−11.9)	0.257 (−10.0)	0.334 (−11.6)	0.396 (−11.3)
	A2	0.749 (−0.7)	0.506 (+1.6)	0.375 (+0.2)	0.442 (−0.7)	0.350 (−0.7)	0.448 (−0.2)	0.501 (−0.9)
	A3	0.769 (+1.3)	0.577 (+8.7)	0.451 (+7.9)	0.353 (−9.5)	0.262 (−9.5)	0.357 (−9.3)	0.413 (−9.6)
GPT-5.6 terra	A0	0.756	0.404	0.295	0.544	0.479	0.566	0.611
	A1	0.709 (−4.7)	0.542 (+13.8)	0.372 (+7.7)	0.411 (−13.2)	0.338 (−14.1)	0.422 (−14.4)	0.476 (−13.5)
	A2	0.708 (−4.8)	0.420 (+1.6)	0.297 (+0.2)	0.490 (−5.4)	0.453 (−2.7)	0.520 (−4.6)	0.580 (−3.0)
	A3	0.703 (−5.3)	0.569 (+16.4)	0.360 (+6.5)	0.429 (−11.5)	0.353 (−12.7)	0.444 (−12.2)	0.489 (−12.1)
Qwen3-VL-32B	A0	0.673	0.347	0.196	0.637	0.550	0.665	0.709
	A1	0.673 (+0.0)	0.250 (−9.6)	0.104 (−9.3)	0.823 (+18.6)	0.631 (+8.1)	0.831 (+16.6)	0.848 (+13.9)
	A2	0.551 (−12.2)	0.347 (+0.0)	0.089 (−10.8)	0.807 (+17.1)	0.661 (+11.1)	0.817 (+15.2)	0.841 (+13.2)
	A3	0.632 (−4.1)	0.305 (−4.2)	0.145 (−5.1)	0.683 (+4.6)	0.499 (−5.1)	0.700 (+3.5)	0.772 (+6.3)
Qwen3-VL-8B	A0	0.665	0.283	0.154	0.721	0.602	0.728	0.770
	A1	0.665 (+0.0)	0.193 (−9.0)	0.059 (−9.5)	0.900 (+17.9)	0.686 (+8.4)	0.903 (+17.4)	0.913 (+14.3)
	A2	0.565 (−10.0)	0.283 (+0.0)	0.051 (−10.3)	0.895 (+17.4)	0.734 (+13.2)	0.901 (+17.3)	0.912 (+14.2)
	A3	0.613 (−5.2)	0.259 (−2.4)	0.077 (−7.7)	0.833 (+11.2)	0.617 (+1.5)	0.836 (+10.8)	0.875 (+10.5)
Qwen3-VL-2B	A0	0.495	0.333	0.102	0.720	0.528	0.749	0.794
	A1	0.495 (+0.0)	0.251 (−8.2)	0.038 (−6.4)	0.895 (+17.6)	0.575 (+4.7)	0.907 (+15.8)	0.923 (+12.9)
	A2	0.447 (−4.9)	0.333 (+0.0)	0.043 (−5.9)	0.836 (+11.6)	0.595 (+6.6)	0.871 (+12.3)	0.904 (+11.0)
	A3	0.582 (+8.7)	0.234 (−9.9)	0.032 (−7.0)	0.923 (+20.3)	0.445 (−8.3)	0.929 (+18.0)	0.945 (+15.1)
SpatioLM	A0	0.511	0.235	0.079	0.754	0.623	0.793	0.848
	A1	0.511 (+0.0)	0.160 (−7.5)	0.010 (−6.9)	0.975 (+22.1)	0.676 (+5.3)	0.980 (+18.6)	0.983 (+13.5)
	A2	0.422 (−8.9)	0.235 (+0.0)	0.013 (−6.6)	0.964 (+21.0)	0.742 (+11.9)	0.968 (+17.4)	0.973 (+12.5)
	A3	0.457 (−5.4)	0.215 (−2.0)	0.011 (−6.8)	0.953 (+19.8)	0.575 (−4.8)	0.960 (+16.6)	0.977 (+12.9)
InternVL3.5-8B	A0	0.523	0.284	0.081	0.781	0.639	0.818	0.847
	A1	0.523 (+0.0)	0.239 (−4.5)	0.044 (−3.7)	0.893 (+11.2)	0.685 (+4.6)	0.913 (+9.5)	0.919 (+7.2)
	A2	0.473 (−5.0)	0.284 (+0.0)	0.033 (−4.9)	0.914 (+13.2)	0.697 (+5.8)	0.926 (+10.8)	0.934 (+8.7)
	A3	0.542 (+1.9)	0.230 (−5.4)	0.033 (−4.8)	0.917 (+13.6)	0.497 (−14.2)	0.922 (+10.4)	0.941 (+9.4)
SpatialRGPT-8B	A0	0.472	0.248	0.040	0.853	0.716	0.892	0.917
	A1	0.472 (+0.0)	0.192 (−5.6)	0.003 (−3.7)	0.993 (+14.0)	0.729 (+1.4)	0.995 (+10.3)	0.995 (+7.8)
	A2	0.427 (−4.5)	0.248 (+0.0)	0.006 (−3.4)	0.985 (+13.2)	0.764 (+4.9)	0.989 (+9.7)	0.989 (+7.2)
	A3	0.548 (+7.6)	0.199 (−4.9)	0.031 (−0.9)	0.934 (+8.2)	0.434 (−28.2)	0.940 (+4.8)	0.946 (+2.9)

- **Language-prior substitution.** The counterfactual answer equals the image-free text-only answer even though it matches neither image. This is the signature that LBR counts.
- **Architecture-dependent read-out.** The spatially specialised and the general models can give opposite counterfactual answers on the same pair, so the appendix reports the vote distribution across models rather than a single averaged answer.
- **Cross-family agreement on the failure, disagreement on the wrong option.** When several models fail the same pair they frequently land on different wrong options, so a black-box bias rate cannot by itself tell whether they share a failure mechanism. That is the question the diagnostic appendix answers at the level of individual layers.

C.3 Error Cases Analysis

We call a pair a *consensus failure* when all 8 models answer the original image correctly and all 8 then commit a bias error on the counterfactual image; such pairs are hard for every architecture at once. Tab. 7 breaks the 81 consensus failures out by spatial and reasoning category.

The 81 consensus failures spread across the question types rather than concentrating in one: every spatial and reasoning category contributes, so the benchmark is not accidentally testing a single spatial relation. The largest shares fall in the categories that require comparing two objects whose depth or 3D layout must be inferred from a single image (geometric structure and physical–geometric reasoning), whereas explicit left/right or above/below position questions are the easiest, since the model can often read the answer off the pixel layout without recovering depth. A model that reaches a high original-image accuracy can therefore still fail the counterfactual partner for a reason that no accuracy number exposes, which the diagnostic appendix examines directly.

Table 4: **Full ablation-B σ sweep across the capability axis.** The target region (counterfactual box grown by 10%) stays sharp while the background is progressively blurred with Gaussian σ. σ=0 is the raw image (identical pixels to A0). Values in parentheses are percentage-point changes from the row marked σ=0 (raw image) for the same model; red = increase, green = decrease.

Model	Cond.	Acc^{O} ↑	Acc^{CF} ↑	pAcc ↑	PRR ↓	LBR ↓	BR ↓	CER ↓
Gemini 3.8 Flash	σ=0 (raw)	0.756	0.490	0.372	0.448	0.357	0.450	0.509
	σ=4	0.758 (+0.2)	0.518 (+2.9)	0.410 (+3.7)	0.408 (−4.0)	0.319 (−3.8)	0.411 (−3.9)	0.460 (−4.9)
	σ=8	0.772 (+1.6)	0.525 (+3.5)	0.419 (+4.7)	0.411 (−3.8)	0.330 (−2.7)	0.418 (−3.3)	0.457 (−5.2)
	σ=12	0.762 (+0.6)	0.506 (+1.7)	0.400 (+2.7)	0.421 (−2.7)	0.331 (−2.6)	0.430 (−2.0)	0.476 (−3.3)
	σ=15	0.759 (+0.3)	0.506 (+1.7)	0.403 (+3.0)	0.415 (−3.4)	0.340 (−1.7)	0.422 (−2.9)	0.468 (−4.1)
GPT-5.6 terra	σ=0 (raw)	0.756	0.404	0.295	0.544	0.479	0.566	0.611
	σ=4	0.742 (−1.4)	0.440 (+3.6)	0.299 (+0.4)	0.542 (−0.2)	0.481 (+0.1)	0.555 (−1.1)	0.596 (−1.4)
	σ=8	0.759 (+0.3)	0.433 (+2.9)	0.301 (+0.6)	0.563 (+1.9)	0.504 (+2.4)	0.575 (+0.9)	0.604 (−0.7)
	σ=12	0.756 (+0.0)	0.433 (+2.9)	0.297 (+0.2)	0.562 (+1.8)	0.513 (+3.4)	0.576 (+1.0)	0.607 (−0.3)
	σ=15	0.743 (−1.3)	0.448 (+4.4)	0.313 (+1.8)	0.533 (−1.1)	0.484 (+0.4)	0.546 (−2.0)	0.574 (−3.6)
Qwen3-VL-32B	σ=0 (raw)	0.673	0.347	0.196	0.637	0.550	0.665	0.709
	σ=4	0.681 (+0.8)	0.384 (+3.7)	0.238 (+4.1)	0.590 (−4.6)	0.508 (−4.2)	0.610 (−5.5)	0.653 (−5.6)
	σ=8	0.673 (+0.0)	0.378 (+3.1)	0.239 (+4.3)	0.572 (−6.4)	0.513 (−3.7)	0.607 (−5.8)	0.646 (−6.3)
	σ=12	0.657 (−1.6)	0.382 (+3.5)	0.237 (+4.1)	0.566 (−7.1)	0.495 (−5.5)	0.591 (−7.4)	0.641 (−6.8)
	σ=15	0.657 (−1.6)	0.381 (+3.4)	0.234 (+3.8)	0.569 (−6.7)	0.499 (−5.1)	0.599 (−6.7)	0.645 (−6.4)
Qwen3-VL-8B	σ=0 (raw)	0.665	0.283	0.154	0.721	0.602	0.728	0.770
	σ=4	0.660 (−0.5)	0.313 (+3.0)	0.191 (+3.7)	0.655 (−6.6)	0.579 (−2.3)	0.666 (−6.3)	0.712 (−5.9)
	σ=8	0.648 (−1.7)	0.313 (+3.0)	0.187 (+3.3)	0.651 (−7.0)	0.575 (−2.8)	0.666 (−6.3)	0.714 (−5.6)
	σ=12	0.647 (−1.8)	0.317 (+3.4)	0.193 (+3.9)	0.639 (−8.2)	0.571 (−3.1)	0.654 (−7.4)	0.703 (−6.7)
	σ=15	0.631 (−3.4)	0.321 (+3.8)	0.193 (+3.9)	0.632 (−8.9)	0.568 (−3.4)	0.648 (−8.1)	0.696 (−7.5)
Qwen3-VL-2B	σ=0 (raw)	0.495	0.333	0.102	0.720	0.528	0.749	0.794
	σ=4	0.487 (−0.8)	0.360 (+2.7)	0.134 (+3.2)	0.643 (−7.7)	0.519 (−0.9)	0.673 (−7.6)	0.725 (−6.9)
	σ=8	0.486 (−0.9)	0.352 (+1.9)	0.129 (+2.7)	0.654 (−6.6)	0.523 (−0.5)	0.685 (−6.4)	0.734 (−5.9)
	σ=12	0.487 (−0.8)	0.341 (+0.8)	0.115 (+1.3)	0.674 (−4.6)	0.545 (+1.7)	0.715 (−3.4)	0.764 (−3.0)
	σ=15	0.493 (−0.2)	0.343 (+1.0)	0.126 (+2.4)	0.650 (−7.0)	0.527 (−0.1)	0.694 (−5.5)	0.744 (−5.0)
SpatioLM	σ=0 (raw)	0.511	0.235	0.079	0.754	0.623	0.793	0.848
	σ=4	0.526 (+1.4)	0.234 (−0.1)	0.082 (+0.3)	0.734 (−2.1)	0.601 (−2.2)	0.766 (−2.8)	0.846 (−0.2)
	σ=8	0.516 (+0.5)	0.238 (+0.2)	0.080 (+0.1)	0.733 (−2.1)	0.598 (−2.4)	0.766 (−2.8)	0.848 (−0.0)
	σ=12	0.518 (+0.6)	0.250 (+1.5)	0.096 (+1.8)	0.705 (−5.0)	0.584 (−3.9)	0.733 (−6.1)	0.816 (−3.2)
	σ=15	0.513 (+0.2)	0.250 (+1.4)	0.096 (+1.7)	0.694 (−6.0)	0.588 (−3.5)	0.727 (−6.6)	0.816 (−3.2)
InternVL3.5-8B	σ=0 (raw)	0.523	0.284	0.081	0.781	0.639	0.818	0.847
	σ=4	0.557 (+3.3)	0.313 (+2.9)	0.119 (+3.7)	0.713 (−6.9)	0.589 (−5.0)	0.757 (−6.1)	0.789 (−5.8)
	σ=8	0.547 (+2.4)	0.317 (+3.3)	0.123 (+4.1)	0.703 (−7.8)	0.576 (−6.3)	0.735 (−8.3)	0.778 (−6.9)
	σ=12	0.544 (+2.1)	0.322 (+3.8)	0.125 (+4.4)	0.694 (−8.7)	0.574 (−6.6)	0.731 (−8.7)	0.772 (−7.5)
	σ=15	0.545 (+2.2)	0.326 (+4.2)	0.131 (+4.9)	0.685 (−9.7)	0.566 (−7.3)	0.721 (−9.7)	0.762 (−8.5)
SpatialRGPT-8B	σ=0 (raw)	0.472	0.248	0.040	0.853	0.716	0.892	0.917
	σ=4	0.469 (−0.3)	0.265 (+1.7)	0.056 (+1.6)	0.801 (−5.2)	0.707 (−0.9)	0.845 (−4.7)	0.882 (−3.5)
	σ=8	0.458 (−1.4)	0.262 (+1.4)	0.050 (+1.0)	0.808 (−4.5)	0.723 (+0.7)	0.845 (−4.7)	0.892 (−2.5)
	σ=12	0.446 (−2.6)	0.260 (+1.2)	0.050 (+1.0)	0.797 (−5.5)	0.742 (+2.6)	0.844 (−4.8)	0.889 (−2.8)
	σ=15	0.440 (−3.2)	0.256 (+0.8)	0.049 (+1.0)	0.797 (−5.6)	0.744 (+2.8)	0.848 (−4.4)	0.889 (−2.8)

D More Experimental Details

D.1 Computation Details

All experiments are inference-only; no model is trained or fine-tuned. The local models run on a single shared node with four RTX 4090 GPUs and consume about 138.3 GPU-hours; the proprietary models are queried through an API and account for about 90,288 calls, which use no local GPU. The diagnostic sweep adds a separate 1254-pair forward+backward pass on Qwen3-VL-8B, sharded over two GPUs, for about 5.3 GPU-hours.

D.2 Background Suppression Details

To examine whether irrelevant visual context contributes to spatial bias, we construct four background configurations, as illustrated in Fig. 9.

B0: Original Image. B0 uses the unmodified image and serves as the reference configuration.

B1: Object-centered Background Suppression. We use SAM 3 (Carion et al., 2026) with the annotated target region to obtain a foreground mask. The target object is preserved, while the surrounding background is progressively blurred with increasing distance from the target, producing an object-centered spatially varying suppression pattern. We evaluate $\sigma \in \{4, 8, 15\}$, where σ denotes

Table 5: **Full ablation-B falsification and mechanism controls.** B2 blurs the whole image uniformly (no sharp target region); B3 leaves every pixel unchanged and only draws a red box around the target region (pure spatial hint). B0 is the raw-image baseline. Values in parentheses are percentage-point changes from the row marked σ=0 (raw image) for the same model; red = increase, green = decrease.

Model	Cond.	Acc^{O} ↑	Acc^{CF} ↑	pAcc ↑	PRR ↓	LBR ↓	BR ↓	CER ↓
Gemini 3.8 Flash	B0 (raw)	0.756	0.490	0.372	0.448	0.357	0.450	0.509
	B2 uniform σ=8	0.726 (−3.0)	0.380 (−10.9)	0.265 (−10.8)	0.583 (+13.5)	0.471 (+11.4)	0.588 (+13.8)	0.637 (+12.8)
	B3 bbox hint	0.780 (+2.4)	0.541 (+5.2)	0.423 (+5.1)	0.410 (−3.8)	0.326 (−3.1)	0.417 (−3.3)	0.456 (−5.3)
GPT-5.6 terra	B0 (raw)	0.756	0.404	0.295	0.544	0.479	0.566	0.611
	B2 uniform σ=8	0.659 (−9.7)	0.304 (−10.0)	0.143 (−15.2)	0.735 (+19.1)	0.648 (+16.8)	0.748 (+18.2)	0.785 (+17.4)
	B3 bbox hint	0.718 (−3.8)	0.549 (+14.5)	0.372 (+7.7)	0.427 (−11.7)	0.348 (−13.1)	0.437 (−12.9)	0.482 (−12.9)
Qwen3-VL-32B	B0 (raw)	0.673	0.347	0.196	0.637	0.550	0.665	0.709
	B2 uniform σ=8	0.612 (−6.1)	0.230 (−11.6)	0.092 (−10.4)	0.782 (+14.5)	0.680 (+13.0)	0.812 (+14.7)	0.852 (+14.3)
	B3 bbox hint	0.701 (+2.8)	0.423 (+7.7)	0.267 (+7.1)	0.540 (−9.6)	0.478 (−7.2)	0.578 (−8.7)	0.620 (−8.9)
Qwen3-VL-8B	B0 (raw)	0.665	0.283	0.154	0.721	0.602	0.728	0.770
	B2 uniform σ=8	0.537 (−12.8)	0.194 (−8.9)	0.076 (−7.8)	0.793 (+7.2)	0.662 (+6.0)	0.810 (+8.1)	0.860 (+9.0)
	B3 bbox hint	0.664 (−0.1)	0.352 (+6.9)	0.219 (+6.5)	0.619 (−10.3)	0.534 (−6.8)	0.626 (−10.3)	0.671 (−9.9)
Qwen3-VL-2B	B0 (raw)	0.495	0.333	0.102	0.720	0.528	0.749	0.794
	B2 uniform σ=8	0.415 (−8.0)	0.283 (−5.0)	0.038 (−6.4)	0.831 (+11.1)	0.637 (+10.9)	0.862 (+11.3)	0.908 (+11.4)
	B3 bbox hint	0.485 (−1.0)	0.383 (+4.9)	0.138 (+3.6)	0.632 (−8.8)	0.507 (−2.2)	0.669 (−7.9)	0.715 (−7.8)
SpatioLM	B0 (raw)	0.511	0.235	0.079	0.754	0.623	0.793	0.848
	B2 uniform σ=8	0.482 (−3.0)	0.194 (−4.1)	0.041 (−3.7)	0.832 (+7.8)	0.669 (+4.7)	0.849 (+5.5)	0.917 (+6.9)
	B3 bbox hint	0.484 (−2.7)	0.238 (+0.3)	0.084 (+0.5)	0.721 (−3.3)	0.607 (−1.6)	0.752 (−4.1)	0.828 (−2.0)
InternVL3.5-8B	B0 (raw)	0.523	0.284	0.081	0.781	0.639	0.818	0.847
	B2 uniform σ=8	0.478 (−4.5)	0.234 (−4.9)	0.038 (−4.3)	0.869 (+8.8)	0.705 (+6.6)	0.888 (+7.0)	0.923 (+7.6)
	B3 bbox hint	0.503 (−2.0)	0.329 (+4.5)	0.116 (+3.4)	0.693 (−8.8)	0.618 (−2.1)	0.733 (−8.5)	0.773 (−7.4)
SpatialRGPT-8B	B0 (raw)	0.472	0.248	0.040	0.853	0.716	0.892	0.917
	B2 uniform σ=8	0.432 (−4.0)	0.212 (−3.6)	0.013 (−2.7)	0.941 (+8.8)	0.793 (+7.7)	0.959 (+6.8)	0.972 (+5.5)
	B3 bbox hint	0.466 (−0.6)	0.275 (+2.7)	0.057 (+1.7)	0.796 (−5.7)	0.703 (−1.2)	0.844 (−4.8)	0.880 (−3.7)

Table 6: **Representative qualitative cases.** One pair per spatial category, chosen so that the eight models disagree as much as possible on the counterfactual answer, among pairs where at least one model commits a bias error. $y^{\text{O}}/y^{\text{CF}}$ are the ground-truth options; the next column lists how the eight models voted on CF, so a mass on y^{O} is a retained prior. "#bias/8" counts how many models bias-fail on the pair.

ID	Question	y^{O}	y^{CF}	CF answers (8 models)	#bias/8
DSCN_021	For pair DSCN_021, comparing A (the prominent truck in the lower-left region) with B (the prominent c…	A	C	A×5, B×1, C×1, D×1	5/8
GSSC_008	In the upper-middle market parking strip, what overall pattern is formed by the five selected parked v…	C	D	A×1, B×2, C×4, D×1	5/8
SAP_048	For pair SAP_048, what spatial action is visibly depicted by A, the prominent pedestrian in the cente…	B	D	A×1, B×5, C×1, D×1	5/8
SPC_014	In the central through lane of the multi-lane bridge road, is there a single vehicle traveling in the …	D	C	A×1, B×1, C×1, D×5	5/8
SRS_050	For pair SRS_050, what is the image-plane relationship of A (the prominent motorcycle in the lower-le…	B	D	A×1, B×5, C×1, D×1	6/8
VOCN_009	In which diagonal image direction is the physical front of the lower-left red-and-white city bus orien…	C	D	A×1, B×2, C×4, D×1	4/8

the standard deviation of the Gaussian blur. This design suppresses increasingly distant context while retaining task-relevant visual evidence.

B2: Uniform Full-image Blurring. As a control, Gaussian blur with $\sigma = 8$ is applied uniformly to the entire image, including both target and background regions. Unlike B1, this setting does

Table 7: **Error-case analysis by question category.** A pair is a *consensus failure* when all eight models answer the original image correctly and all eight then commit a bias error on the counterfactual image. The consensus set is therefore a lower bound on the pairs that are hard for every architecture at once, and it is not weighted by any single model.

Category	Pairs	Consensus failures
Spatial category		
Depth & scale	187	10
Geometric structure	202	20
Spatial action	206	16
Spatial position	234	2
Spatial relationship	215	12
Viewpoint & orientation	210	21
Reasoning category		
Cultural & behavioural norms	331	15
Egocentric–allocentric	309	14
Physical–geometric	308	39
Semantic co-occurrence	306	13
All pairs	1254	81

not preserve the target region and therefore tests whether the effects observed under B1 arise from selective background suppression rather than generic image degradation.

B3: Bounding-box Highlighting. We additionally provide the annotated target bounding box as an explicit visual cue while leaving the image content otherwise unchanged. This configuration evaluates whether directly guiding the model toward task-relevant regions can improve visual grounding without removing background information.

Together, B0–B3 isolate the effects of *irrelevant-context suppression*, *generic visual degradation*, and *explicit visual guidance*. All configurations preserve the original questions, candidate answers, and ground-truth labels, ensuring that performance differences arise only from changes to the visual input.

D.3 Bias Mitigation Details

This section provides the complete implementation details of the four mitigation strategies introduced in Sec. 5. For the three inference-time strategies, model parameters remain unchanged and only the prompting context is modified relative to the original P0 evaluation. For supervised fine-tuning (SFT), we explicitly expose the model to paired original–counterfactual supervision while preserving the same held-out evaluation protocol.

D.3.1 Prompting Strategies

Zero-shot Linguistic Intervention. This strategy directly tests whether spatial bias can be mitigated through explicit instruction alone. We augment the baseline evaluation prompt with two short constraints that encourage the model to prioritize the currently observed visual evidence and suppress reliance on learned priors. No examples, reasoning demonstrations, or parameter updates are introduced.

P1: Zero-shot Linguistic Intervention

Base your judgment solely on direct visual evidence in the provided image. Do not use your prior knowledge to choose the answer.

Chain-of-Thought Incorporation. To reduce shortcut reasoning, we introduce a structured visual reasoning procedure before answer selection. The prompt explicitly requires the model to identify the target, determine the relevant reference frame and spatial evidence, and verify the selected option against visible contradictions. This intervention tests whether encouraging explicit evidence-grounded reasoning can reduce prior-driven spatial decisions without modifying model parameters.

P2: Chain-of-Thought Incorporation

Work through the visual task step by step before choosing the final option:
1. Locate the exact target object(s) in the TARGET image, using the question's descriptors. State only visible identifying cues; do not invent hidden details.
2. Specify the reference frame and decision rule in the question. Keep image-plane position/projected size distinct from scene depth/physical size and object-centered orientation. Describe the visible relation or property that answers this question.
3. Compare that observation with the options. Check the best competing option for a visible contradiction. Prefer what is depicted over what is normally expected. Select an insufficient-visibility option only if the target evidence supports it.
Output four lines. Keep each of the first three lines to one short sentence and keep the total explanation under 100 words. Do not mention an answer letter until the final line. Use this format:
TARGET: <visible target identification>
EVIDENCE: <reference frame and observed spatial evidence>
CHECK: <brief option comparison grounded in that evidence>
ANSWER=<one uppercase letter A, B, C, or D>
The final line must be the only ANSWER= marker. Do not output any text after it.

Few-shot Prompting. We further test whether counterfactual behavior can be induced through in-context demonstrations. Three representative original–counterfactual examples are provided before the target query. Each demonstration keeps the question and answer options fixed while showing that changes in visual evidence may alter the correct answer. The demonstrations are explicitly presented as independent examples to discourage answer-pattern copying or transfer of scene-specific spatial facts.

P3: Few-shot Prompting

You will first see three worked examples. Each example contains an ORIGINAL image and a COUNTERFACTUAL (CF) image, the same question and options, and the correct option for each image. The examples show that changing visible evidence can change the correct answer even when the question and options stay the same. Learn to judge the image actually supplied, rather than repeating an earlier answer or choosing what would normally be expected. A new answer is allowed to change, but do not force a change merely because an image looks unusual. The examples are independent scenes. Their answer letters are not a pattern to copy. After the examples, answer only the TARGET question using the TARGET image. Never transfer an object or a spatial fact from an example to the target.
EXAMPLE 1 — ORIGINAL image: ...
EXAMPLE 1 — CF image: ...
EXAMPLE 1 question (applies to both images): ...
EXAMPLE 1 options: A. ... B. ... C. ... D. ...
Correct option for ORIGINAL: ...
Correct option for CF: ...
The same question has different correct options because the visible evidence changed.
END OF EXAMPLE 1.
EXAMPLE 2 ...
EXAMPLE 3 ...

D.3.2 Supervised Fine-tuning Configuration

Data Split and Supervision. We investigate parameter-level mitigation by fine-tuning each model on paired original–counterfactual supervision. The dataset is split into 877 training, 126 validation, and 251 test image pairs, corresponding approximately to a 7:1:2 ratio. Each original–counterfactual pair is kept within the same split. Samples sharing identical image hashes, common source images, or identifiable video sequences are also grouped to minimize source leakage. The same partition is used for all evaluated models.

Each training pair is converted into two independent image–question–answer examples, resulting in 1,754 training instances. The original and counterfactual images share the same question and answer options but are supervised using their respective ground-truth answers. This design explicitly teaches the model that identical linguistic inputs may require different predictions when the visual evidence changes.

Training Objective. We optimize autoregressive cross-entropy over the assistant response only. The target response follows the format `ANSWER=<letter>` together with its termination token, while prompt and image tokens are masked from the loss. No text-only samples or intermediate reasoning traces are included during fine-tuning, allowing us to isolate the effect of paired visual supervision.

Parameter-efficient Fine-tuning. We adopt LoRA-based parameter-efficient fine-tuning while keeping the original model parameters frozen. LoRA adapters are inserted into the attention and feed-forward projections of the language decoder. The main training configuration is summarized below.

Model Selection and Evaluation. For each model, we select the epoch checkpoint with the lowest validation response-token cross-entropy. At test time, the fine-tuned model uses the same P0 evaluation prompt and image preprocessing as the unmodified baseline. All mitigation strategies are evaluated on the same held-out test set to ensure a controlled comparison.

Table 8: SFT configuration for spatial-bias mitigation.

Configuration	Value
Training pairs	877
Validation pairs	126
Test pairs	251
Training instances	1,754
LoRA rank	16
LoRA scaling factor	32
LoRA dropout	0.05
Optimizer	AdamW
Learning rate	5×10^{-5}
Weight decay	0.01
Effective batch size	16
Training epochs	3
LR schedule	Cosine
Warmup steps	17

E Bias Diagnostic Framework

Positioning and relation to the black-box pipeline. Sec. 4 classifies a counterfactual failure *behaviourally*, using only the O/CF/T responses of the same benchmark: whether the answer is retained from the original image, whether it coincides with the image-free text-only prediction, or whether it matches neither. That pipeline answers *what* the failure looks like, but it cannot say *where* in the forward pass the spatial evidence stops to matter. This appendix complements it with a *white-box* diagnostic that localizes each failure to one of three measurement stages of VLM inference: *(L1)* the visual encoder, *(L2)* the multimodal interaction that writes visual information into the answer position, and *(L3)* the language-side decision. Throughout, we inherit the sample universe and the bias definitions of the main Tab., so that the diagnostic is applied to the samples the main pipeline calls a counterfactual failure (Sec. 4, Stage 2), and add no new ground-truth assumptions of our own. Of those failures, the subset whose region is localizable receives a layer-wise label; the remainder is reported as `out_of_scope` rather than attributed to a layer, and the split is given in Tab. 9.

E.1 Fine-grained Spatial Error Diagnostic Framework

E.1.1 Scope and Pre-diagnostic Gate Control

Not every O/CF pair admits a physically meaningful paired measurement, because the two images may not be token-wise comparable. We therefore apply two gates *before* any layer-wise quantity is computed; a pair that fails either gate is labelled `out_of_scope` and excluded from all paired analyses (per-region statistics are not reported for it).

- **Grid gate (token alignment).** Let $\mathrm{Grid}(I)$ be the visual feature grid. We require $\mathrm{Grid}(I^O) = \mathrm{Grid}(I^{CF})$, i.e. the pre-merge and post-merge heights and widths of the two images are identical. When the two images have different aspect ratios or resolutions, visual token k of O and token k of CF no longer refer to the same location, and every *paired* quantity below (Δ features, attention share, write-in flow) silently misaligns.

Table 9: **Diagnostic scope.** Measurability and failure are independent axes, not a chain. The grid gate removes the pairs whose token grids disagree (13 mismatches and 1 evaluation error); the locality gate then splits the rest into the in-scope pool, on which the chain runs, and the pairs reported as `out_of_scope`. A *failure* is a pair with O correct and CF wrong, inherited from the main table. The in-scope pool is the basis of the real-vs-null validation.

	In-scope (measurable)	**Out of scope**	**Total**
O correct / CF wrong (failures)	514	119	633
Otherwise	489	118	607
Total	**1003**	**237**	**1240**

- **Locality gate (bounded intervention).** Let $\mathcal{T}_{\text{edit}}$ be the set of visual tokens covered by the edited region. We require $|\mathcal{T}_{\text{edit}}| \leq 0.5\,|\mathcal{T}_{\text{vis}}|$. Ablating or comparing regions that cover most of the image does not isolate a local spatial edit and risks pushing the model far outside the input distribution it was evaluated on.

Together the two gates fix the universe on which the diagnostic is defined: the grid gate removes the images whose token counts disagree by construction, and the locality gate partitions the remainder into an *in-scope* pool, on which the chain runs, and the `out_of_scope` rest. On this benchmark the area-ratio test is the only reason a pair falls in the latter. Two axes are consequently at work, and they are independent: *measurability* (does the pair admit a localizable paired measurement?) and *failure* (does the model keep the wrong answer on the counterfactual image?). Tab. 9 crosses them, and the distinction governs denominators throughout the appendix: the region-level validation against the null uses the in-scope pool, since random blocks are localizable by construction and the two sides must be comparable, whereas the attribution of failures uses only those failures that are also measurable.

Protocol alignment. All internal quantities are measured with the *identical* prompt template, answer prefix, and image resolution (no downscaling) used by the main Tab., and the one-step decision is read from the A/B/C/D next-token logits. Because a one-step argmax and free generation can disagree on borderline samples, the *gate and language labels* are taken from the main Tab. (matched by sample id) while the *internal measurements* come from our forward pass; the shared prompt and resolution guarantee that the answer-position hidden state is the same object in both. Any residual disagreement is thus a labelling difference, never a measurement difference. Cross-checks of the shared state (image grid, visual token count, and generation-based prediction) are recorded in the diagnostic run log; the grid-gate counts in Tab. 9 are the part of that check that bears on scope.

E.1.2 Region Construction

For each eligible pair we define one *edit* region and, when a second annotated object exists, one *reference* region:

- **Target region** $\mathcal{R}_{\text{tgt}}$: the union of the original and counterfactual annotation boxes, grown by 10% and mapped to the feature grid; a cell is included if it overlaps the box by at least 0.5.
- **Reference region** $\mathcal{R}_{\text{ref}}$: the annotated reference-object box grown by 10%, with all target cells removed. Reference regions exist only for comparison-style questions.

We write $\mathcal{T}_{\text{edit}}$ for the set of post-merge visual tokens covered by $\mathcal{R}_{\text{tgt}}$, and $\mathcal{T}_{\text{vis}}$ for the full set of post-merge visual tokens. Every layer-wise quantity below is indexed by these token sets, since attention rows and gradients are defined per token rather than per cell of the box. Because the target region is the union of the two boxes, it is a *superset* of the pixels that actually changed; this dilutes the measured Δ by a factor that cannot be removed without per-pixel annotation. We therefore always report the target region together with a *same-size random block* null (Sec. E.1.4), which absorbs the dilution into the baseline instead of hiding it.

E.1.3 Layer-wise Quantitative Metrics

L1 — visual representation layer ("was the change encoded at all?"). Let F^O, F^{CF} be the visual features at the encoder output (pre-merge), restricted to $\mathcal{T}_{\text{edit}}$. We measure two complementary

Table 10: **Instruments of the diagnostic framework.** Each layer contributes one claim type; none of them is derived from the black-box predictions. $\mathcal{T}_{\text{edit}}$ is the token set of the target region and $\mathcal{T}_{\text{vis}}$ the set of all visual tokens.

Stage	Instruments	Computation	Claim type	Null baseline
L1 visual encoder	$d_{\text{cos}}^{\text{vis}},\ d_{\text{rel}}^{\text{vis}}$	forward (features)	representation	same-size random block
L2 interaction	$r_{\text{flow}},\ A_{\text{lift}}$	forward (attention/value)	information routing	same-size random block
L3 decision (sensitivity)	Sal_{norm}	one backward pass	local sensitivity	same-size random block
L3 decision (causal)	effect	mean-ablation, repeated forward	causal contribution	ablation on random blocks

distances in that space:

$$d_{\text{cos}}^{\text{vis}} = 1 - \cos\big(\bar{F}^{O}, \bar{F}^{CF}\big)\,, \qquad d_{\text{rel}}^{\text{vis}} = \frac{\|F^{CF} - F^{O}\|_2}{\|F^{O}\|_2 + \epsilon}, \tag{1}$$

where $\bar{F}$ denotes the mean token. $d_{\text{cos}}^{\text{vis}}$ is scale-invariant and bounded, while $d_{\text{rel}}^{\text{vis}}$ retains amplitude sensitivity; the layer-1 statistic is $\phi_1 = \max\big(d_{\text{cos}}^{\text{vis}}, d_{\text{rel}}^{\text{vis}}\big)$.

L2 — multimodal interaction layer ("did the change reach the answer position?"). We interrogate the write-in path from the edit-region tokens to the *answer position* (the last token that emits the option letter). Let $\mathcal{T}_{\text{edit}}$ denote the post-merge tokens of the target region and $\mathcal{T}_{\text{vis}}$ the set of all visual tokens. With $A^{(l,h)}, V^{(l,h)}$ the attention matrix and value tensor of layer l, head h (W_O the output projection), define the edit-region write-in difference and the total visual write-in as

$$\Delta O_{\text{edit}} = \sum_{l,h} W_O^{(l,h)} \big[\,(A^{CF}V^{CF})_{:,\mathcal{T}_{\text{edit}}} - (A^{O}V^{O})_{:,\mathcal{T}_{\text{edit}}}\,\big], \quad O_{\text{all}} = \sum_{l,h} W_O^{(l,h)}\,(A^{O}V^{O})_{:,:}, \tag{2}$$

$$r_{\text{flow}} = \|\Delta O_{\text{edit}}\|_2 \big/ \|O_{\text{all}}\|_2, \qquad A_{\text{lift}} = \tfrac{1}{L}\sum_{l=1}^{L} \frac{\sum_{j\in\mathcal{T}_{\text{edit}}} A^{(l)}_{-1,j}}{\sum_{j\in\mathcal{T}_{\text{vis}}} A^{(l)}_{-1,j}} \Big/ \frac{|\mathcal{T}_{\text{edit}}|}{|\mathcal{T}_{\text{vis}}|}.$$

r_{flow} measures *how much the write-in from the edit region changes* between O and CF (detectability of the edit); A_{lift} measures *how much attention the answer position already pays* to the region relative to its area (reliance on the region). Both are informative on the real edit region (Tab. 13), and they capture distinct failure modes: a change may be visible in the attention but not propagate through the value path, or vice versa.

L3 — logical decision layer ("did the decision use it?"). We use two instruments of different logical strength. First, a *local sensitivity*: a single backward pass gives the mean per-token gradient norm of the counterfactual ground-truth logit $z_{y^{CF}}$ with respect to the edit-region features,

$$\text{Sal}_{\text{norm}} = \frac{1}{|\mathcal{T}_{\text{edit}}|} \sum_{j\in\mathcal{T}_{\text{edit}}} \big\|\partial z_{y^{CF}} / \partial F_j\big\|_2. \tag{3}$$

Second, a genuine *causal intervention*: we run the counterfactual forward pass normally to obtain $P^{CF}(\text{correct})$, then replace the edit-region visual features by the mean of the *non-edit* visual tokens of the same image (norm-matched to the original edit-region tokens, so that no out-of-distribution all-zero vector is injected), and re-run:

$$\text{effect} = P^{CF}(\text{correct}) - P^{\text{abl}}(\text{correct}). \tag{4}$$

Sal_{norm} is a first-order statement about *sensitivity*; `effect` is a statement about the *causal contribution* of the region. We keep the two claims strictly separate and label the decision layer with the intervention, not with the gradient. Tab. 10 summarizes the instruments and their claim type.

E.1.4 Null Baseline Construction

To provide a parameter-free baseline without introducing global distributional assumptions, each target edit region $\mathcal{R}_{\text{tgt}}$ is paired with K *same-size* random background blocks drawn from the same image. We maintain two distinct null budgets according to computational cost:

- **Metric nulls** (K_{metric} blocks) supply the sample-specific baseline for representation, routing, and sensitivity metrics (Gates 1–4). They are read directly off the *same* feature, attention, and gradient tensors by permuting token indices, incurring zero additional forward passes.

Algorithm 1 Deterministic layer-wise attribution of a counterfactual failure.

Require: eligible pair (I^O, I^{CF}, q); target region $\mathcal{R}_{\text{tgt}}$ with token set $\mathcal{T}_{\text{edit}}$; budgets $K_{\text{metric}}, K_{\text{abl}}$
Ensure: one label from Tab. 11

if $\text{Grid}(I^O) \neq \text{Grid}(I^{CF})$ or $|\mathcal{T}_{\text{edit}}| > 0.5\,|\mathcal{T}_{\text{vis}}|$ **then**
  return `out_of_scope` ▷ Step 0: region not localizable
end if
draw K_{metric} same-size random blocks; compute their per-metric means $\overline{(\cdot)}^{\text{null}}$
compute $\phi_1, r_{\text{flow}}, A_{\text{lift}}, \text{Sal}_{\text{norm}}$ on $\mathcal{T}_{\text{edit}}$
if ϕ_1 or $\overline{\phi_1}^{\text{null}}$ unavailable **then**
  return `undetermined` ▷ Gate 1 cannot be evaluated
end if
if $\phi_1 \leq \overline{\phi_1}^{\text{null}}$ **then**
  return `feature_collapse` ▷ the encoder did not encode the edit
end if
if $r_{\text{flow}}, \overline{r_{\text{flow}}}^{\text{null}}$ available **and** $r_{\text{flow}} \leq \overline{r_{\text{flow}}}^{\text{null}}$ **then**
  return `alignment_fail` ▷ the edit is not routed to the answer position
end if
if $A_{\text{lift}}, \overline{A_{\text{lift}}}^{\text{null}}$ available **and** $A_{\text{lift}} \leq \overline{A_{\text{lift}}}^{\text{null}}$ **then**
  return `attention_failure` ▷ the answer position ignores the region
end if
if Sal_{norm} or $\overline{\text{Sal}_{\text{norm}}}^{\text{null}}$ unavailable **then**
  return `undetermined` ▷ Gate 4 cannot be evaluated (no gradient)
end if
if $\text{Sal}_{\text{norm}} \leq \overline{\text{Sal}_{\text{norm}}}^{\text{null}}$ **then**
  return `language_prior` ▷ the decision does not depend on the region
end if
$\text{effect} \leftarrow P^{CF}(\text{correct}) - P^{\text{abl}}(\text{correct})$ ▷ mean-ablation on $\mathcal{T}_{\text{edit}}$
if effect unavailable **then**
  return `undetermined` ▷ Gate 5 cannot be evaluated
end if
if $\text{effect} > 0$ **then**
  return `reasoning_error` ▷ vision supports the correct answer, but is overridden
else
  return `understanding_error` ▷ vision pushes the decision to the wrong answer
end if

Offline self-check: rerun the chain above on every random block, using the remaining blocks as leave-one-out nulls (Sec. E.1.6).

- **Ablation nulls** ($K_{\text{abl}} \ll K_{\text{metric}}$ blocks) supply the null distribution for the intervention `effect`. Because each block requires an independent forward pass, only a small budget is evaluated for the null self-check. Importantly, this budget affects *only* the baseline verification (Sec. E.1.6) and does not alter the deterministic real-region gates, where Gate 5 relies solely on $\text{sign}(\text{effect})$.

E.1.5 Deterministic Decision Chain

Using the sample-wise null baselines constructed above, the diagnostic applies a deterministic, parameter-free rule independently to each counterfactual failure. Specifically, every metric calculated on the real edit region $\mathcal{R}_{\text{tgt}}$ is compared directly against the empirical mean of the K_{metric} random background blocks from the *same sample* (denoted as $\overline{(\cdot)}^{\text{null}}$).

The decision chain operates as a strict sequence of elimination steps, given as Algorithm 1.

Tab. 11 provides a summary of the full label hierarchy. A pair that fails the Step 0 gates is `out_of_scope` rather than `undetermined`, since its region is by construction not localizable. Within the cascade, a missing statistic is handled in one of two ways:

Table 11: **Diagnostic label set.** Each label names the first elimination step that the sample fails, i.e. the earliest stage at which the edited visual evidence stops influencing the answer.

Label	Stage	Meaning
`feature_collapse`	L1 (Gate 1)	encoder did not encode the edit
`alignment_fail`	L2 (Gate 2)	edit not routed to the answer position
`attention_failure`	L2 (Gate 3)	answer position does not look at the region
`language_prior`	L3 (Gate 4)	decision does not rely on the region
`reasoning_error`	L3 (Gate 5)	vision helps the correct answer; decision overrides it
`understanding_error`	L3 (Gate 5)	vision misleads the decision
`undetermined`	fallback	required statistic unavailable
`out_of_scope`	Step 0	grid or locality gate not satisfied

- **Strict break.** If the Gate 1 statistic, the Gate 4 gradient, or the Gate 5 intervention is unavailable, the chain cannot be completed and the pair is labelled `undetermined`.
- **Soft bypass.** If a Gate 2 or Gate 3 statistic is unavailable, that gate is skipped and the chain continues, so a pair can in principle receive an L3 label without the L2 layer having been tested. The asymmetry is deliberate: an unreadable L2 statistic is treated as no evidence against the L2 explanation, whereas a missing gradient or intervention leaves the elimination chain unfinished. Algorithm 1 encodes both rules.

Note that `language_prior` here is a decision-layer label derived from *local* sensitivity to the edited region; it is distinct from the image-free text-only behavior in the main evaluation, which is recorded as an independent global stratification variable.

E.1.6 Baseline Self-Inspection

Because each metric rule takes the form "$\leq$ null mean", a random non-informative rule would naturally fire on roughly half of the null blocks. We therefore validate the diagnostic by applying the *identical* decision chain to the random background blocks themselves, using a leave-one-out procedure among the K blocks to establish a null label distribution. The true diagnostic conclusion is defined by the *departure* of real edit regions from this background distribution, rather than raw uncalibrated label counts. The comparison is stated at the level of the *region*, not of the failure: a random block has no notion of being correct or wrong, so the null pool cannot be conditioned on failure, and the real side of the comparison is the in-scope pool. Tab. 12 reports both distributions, Tab. 13 shows that every threshold gate fires less often on the real edit region than on a random block of the same size (a conditional ratio below one), and Tab. 14 reports the two quantities that describe the chain as a whole. Two cautions apply when reading the first of these. The ratios of Tab. 12 are *share* ratios — the end state of a cascade — and not gate effectiveness; the cascade removes null blocks and real regions at very different rates, so a late-gate share can point the opposite way from that gate's own conditional ratio, and only Tab. 13 answers the question of whether a gate is informative. And for the labels assigned at Gate 5 the comparison must use the *effect-evaluable* null column, since the remaining blocks cannot be scored there and would otherwise enter the denominator.

E.2 Diagnostic Examples

The three cases below are the exhaustive attributions of a single behavioural signature — a counterfactual failure — rather than three distinct black-box phenomena; their population shares are in Tab. 12. Tab. 15 instantiates every label of the chain on one benchmark pair each, giving the pair's id together with the readings the chain actually judged, so every claim below can be checked against a single row.

Case 1 — visual-level failure (`feature_collapse`): the edit is in the pixels, the answer is unchanged. The sharpest instance is `VOC_007` (Tab. 15): its region changes less than a same-size random block does. Gate 1 fires on the encoder output, before any language-side computation, so the failure precedes the decision. **Mechanism.** The change measured over the annotated region does not exceed what a same-size random block yields, i.e. the edit is never represented — as opposed to

Table 12: **Label distribution: real vs. null.** The null columns come from running the *identical* chain on leave-one-out same-size random blocks. Random blocks are localizable by construction, whereas the *all* pool contains the `out_of_scope` pairs, so the like-for-like comparison is against the *in-scope* columns; the *gated* columns are the failure-conditioned view. The two null columns differ in their denominator: the right one keeps only the blocks whose downstream statistic exists (25% of the pool has no ablation effect), and for the Gate 5 labels (`reasoning_error`, `understanding_error`) it is the appropriate comparison, since the *all* column would divide by blocks that cannot be scored there. Percentages are of each pool; n is given in the last row.

Label	Real (all)	Real (in-scope)	Real (gated)	Real (in-scope ∩ gated)	**Null (all)**	**Null (eff.-eval.)**
`feature_collapse`	11.0%	13.7%	11.7%	14.4%	52.9%	52.9%
`alignment_fail`	10.8%	13.4%	12.3%	15.2%	30.2%	29.9%
`attention_failure`	20.2%	25.0%	20.5%	25.3%	7.9%	8.3%
`language_prior`	3.1%	3.9%	3.6%	4.5%	3.4%	3.2%
`reasoning_error`	19.5%	24.1%	15.8%	19.5%	0.7%	2.8%
`understanding_error`	16.1%	19.9%	17.2%	21.2%	0.7%	2.9%
`out_of_scope`	19.1%	0.0%	18.8%	0.0%	0.0%	0.0%
`undetermined`	0.0%	0.0%	0.0%	0.0%	4.1%	0.0%
n	1240	1003	633	514	36617	9230

Table 13: **Gate effectiveness.** *Conditional* firing rate of each threshold gate: the fraction of the samples that *reach* the gate for which it fires. Raw (unconditional) rates are not comparable across the gates of a cascade — each gate runs only on the survivors of the previous ones — and can even invert the sign of the effect. Real: target region, in-scope $n = 1003$; null: leave-one-out same-size random blocks, $n = 36617$. A ratio below one means the real edit region triggers the gate less often than a random block, i.e. the gate is informative.

Gate	**Statistic**	**Real**	**Null**	**Real/null**
1 (L1)	$\max(d_{\cos}^{\text{vis}}, d_{\text{rel}}^{\text{vis}})$	13.66%	52.89%	0.258
2 (L2)	r_{flow}	15.47%	64.16%	0.241
3 (L2)	A_{lift}	34.29%	46.83%	0.732
4 (L3)	Sal_{norm}	8.11%	38.42%	0.211

represented and then unused, which is the attentional failure of Case 2, and which no saliency map can distinguish from it on its own. The claim is representational, not semantic.

Case 2 — interaction-level bottleneck (`alignment_fail`, `attention_failure`): the edit is encoded, the answer is unchanged. The sharpest instances are `SRP_003` and `DSSC_001` (Tab. 15), whose region is written into and read out of the answer position, respectively, less than a random block is. Gate 1 passes, and Gate 2 or 3 fires. **Mechanism.** For a fixed answer triple a perception failure and a routing failure are observationally equivalent, and Gate 1 eliminates the former; a language-prior account is eliminated by position in the chain, since the diagnosis stops before L3 and therefore reports that the region was not written into the answer position rather than that the decision declined to use it. Gate 2 separates the write from the read. The claim concerns the route, not the content of the evidence.

Case 3 — decision-level misalignment (`language_prior`, `reasoning_error`, `understanding_error`): the edit is encoded and routed, yet the answer is still wrong. The sharpest instances are `SAS_022`, `SRE_031` and `EAT_021` (Tab. 15). **Mechanism.** L3 has two sub-cases. Gate 4 fires when the decision's local sensitivity to the region does not exceed the null, i.e. the answer was formed without depending on the region at all (`language_prior`). If Gate 4 passes, Gate 5 labels by the sign of the mean-ablation effect: removing the region lowering the probability of the correct answer means the evidence supported it and the aggregation discarded it (`reasoning_error`), while raising it means the evidence pushed the decision away (`understanding_error`). The two Gate-5 labels have identical predictions, so only the sign of a causal contribution separates them. Gate 5 is a classifier rather than a threshold and therefore has no null baseline, its margin over the null being the smallest in the chain (Tab. 14).

Table 14: **Whole-chain quantities.** Neither row is a gate's own performance. *Reach* is the product of Gates 1–4 and carries no information about Gate 5; *labels* additionally counts the Gate 4 firings. Gate 5 has no null baseline — its only statistic is the sign split, which is the weakest in the chain and is therefore indicative rather than conclusive.

Quantity	Real	Null (all)	Null (eff.-eval.)
The cascade as a whole (not a gate's own performance)			
Decision-layer **reach** (passes Gates 1–4)	44.07%	5.53%	5.67%
Decision-layer **labels** (Gate 4 fires *or* Gate 5 labels)	47.96%	4.88%	8.89%
Gate 5: a terminal classifier, not a threshold gate			
Sign split $\text{effect} > 0 : \text{effect} < 0$	242 : 199	254 : 264	—
Odds of a positive effect, pos/neg	1.216	0.962	—
n with a usable ablation effect	442	—	523

Table 15: **One representative failure per label.** Each row is a single benchmark pair, identified by its `id`, and reports the *deciding* statistic of that pair together with its leave-one-out null baseline: "real" is the target edit region, "null" the mean over the other random blocks of the same pair. The pair shown for a label is the one whose deciding gate fired by the widest margin within the in-scope ∩ gated pool, so each row is the clearest instance of that failure mode rather than a typical one. Margins are shortfalls from each gate's own null and are therefore comparable *only within a label*: the deciding statistic has different units in every row (a feature distance for Gate 1, a write-in ratio for Gate 2, an attention share for Gate 3, a gradient norm for Gate 4, a probability difference for Gate 5), so the column is not a ranking. Gate 5 labels by the sign of the intervention effect rather than by a threshold test and has no baseline column (shown as "–"); its margin is $|\text{effect}|$. All numbers are produced by `diagnosis/regate.py`; "real", "null" and "margin" are the readings the chain actually judged.

Stage	Label	Pair id	Deciding stat.	Real	Null	Margin (within-label)
L1	`feature_collapse`	`VOC_007`	$\max(d^{\text{vis}}_{\cos}, d^{\text{vis}}_{\text{rel}})$	0.125	0.714	0.589
L2	`alignment_fail`	`SRP_003`	r_{flow}	0.107	0.477	0.370
L2	`attention_failure`	`DSSC_001`	A_{lift}	0.318	2.85	2.54
L3	`language_prior`	`SAS_022`	Sal_{norm}	0.204	0.626	0.422
L3	`reasoning_error`	`SRE_031`	effect	0.833	–	0.833
L3	`understanding_error`	`EAT_021`	effect	-0.951	–	0.951

Caveats. Two properties of the framework limit the strength of its claims and are reported explicitly. *(i) Region precision.* The target region is the union of the two annotation boxes, hence a superset of the changed pixels; L1 magnitudes are diluted accordingly. This cannot be removed without pixel-level annotation and is absorbed by the same-size null baseline. *(ii) Resolution dependence.* Attention- and gradient-based quantities degrade when the region contains only a handful of tokens, so all measurements are taken at the native image resolution (matching the main Tab.); conclusions are stated per model/resolution operating point and are not transferred across resolutions.

F Spatial Bias Benchmark

We construct a paired visual question-answering benchmark to examine whether spatial judgments follow image evidence when familiar spatial regularities are disrupted. Each instance contains an original aerial image I_i^{O}, an edited counterfactual image I_i^{CF}, a shared question q_i, four shared answer options $\mathcal{A}_i$, and image-specific ground-truth labels y_i^{O} and y_i^{CF}:

$$d_i = \big(I_i^{\text{O}}, I_i^{\text{CF}}, q_i, \mathcal{A}_i, y_i^{\text{O}}, y_i^{\text{CF}}, t_i, b_i\big), \tag{5}$$

where t_i denotes the spatial task and b_i denotes the hypothesized source of bias. Holding the question and answer options fixed supports a direct comparison of predictions across the two visual conditions.

The taxonomy crosses six spatial tasks, including spatial position, spatial relationship, spatial action, geometric structure, depth and scale, and viewpoint and orientation, with four bias-reason labels: cultural and behavioral norms, physical-geometric principles, semantic co-occurrence, and egocentric–

allocentric transformation. The final release contains 1,254 pairs (2,508 image files), with all 24 task–reason combinations represented. Individual combinations contain 38–65 pairs (Tab. 16).

Table 16: Final benchmark composition in image pairs. Columns denote cultural/behavioral norms (Cul.), physical-geometric principles (Phy.), semantic co-occurrence (Sem.), and egocentric–allocentric transformation (Ref.).

Spatial task	Cul.	Phy.	Sem.	Ref.	Total
Spatial Position	65	56	59	54	234
Spatial Relationship	58	52	53	52	215
Spatial Action	51	54	51	50	206
Geometric Structure	55	53	46	48	202
Depth and Scale	49	38	49	51	187
Viewpoint and Orientation	53	55	48	54	210
Total	331	308	306	309	1,254

F.1 Dataset Construction Details

F.1.1 Source Imagery

The benchmark is primarily constructed from `VisDrone` (Du et al., 2019), `M3OT` (Nie et al., 2025), `UAVDT` (Du et al., 2018), and `MDMT` (Liu et al., 2023b). Existing annotations are used to localize candidate targets and support target identification during counterfactual construction. Target descriptions rely on attributes that remain valid after editing, while spatial cues affected by the intervention are avoided. Annotations are used only for construction and are not rendered as boxes or markers in the released benchmark images.

F.1.2 Data Construction

Task specification and counterfactual intervention. Each sample is constructed according to a task-specific specification defining the target spatial variable, original and counterfactual states, target scope, question, candidate options, and answer mapping. The counterfactual transformation is formulated as

$$I_i^{\mathrm{CF}} = \mathcal{T}\big(I_i^{\mathrm{O}}; s_i \leftarrow s_i'\big)\,, \tag{6}$$

where s_i denotes the target spatial factor and $\mathcal{T}$ the corresponding editing operation. The intervention may involve a single object, an object–reference relation, or a multi-object configuration. Camera viewpoint, image crop, object identity, and task-irrelevant scene content are preserved as much as possible, while only the spatial factor required by the target task is modified. The six task categories cover absolute position, pairwise spatial relations, spatial actions, geometric structure, depth/scale, and viewpoint/orientation.

Image realization. Counterfactual images are produced through annotation-guided visual editing. The target region is localized from existing annotations, after which the designated object is translated, rotated, or rescaled according to the required spatial intervention. Vacated regions are filled using surrounding visual content, and softened masks are applied to reduce boundary artifacts. For position-related edits, object scale is adjusted when necessary to maintain plausible perspective. All edited images are resized to the original resolution. The overall objective is to modify only the designated spatial factor while preserving non-target scene semantics.

Question construction. Each original–counterfactual pair shares one four-option question. Candidate answers are mutually exclusive, and an explicit reference frame is specified when required. Questions are grounded in the final edited image and avoid wording that reveals which image is counterfactual. We further avoid unsupported assumptions about speed, intent, jurisdiction-specific rules, or metric depth. The intended construction satisfies $y_i^{\mathrm{O}} \neq y_i^{\mathrm{CF}}$. Among the 1,254 retained pairs, 1,251 satisfy this condition, while `SAC_007`, `SAC_009`, and `SPE_004` retain identical answer labels. Answer positions are approximately balanced: the A–D counts are $(308, 316, 314, 316)$ for original images and $(319, 316, 309, 310)$ for counterfactual images.

Quality Control. Quality control verifies target identifiability, successful realization of the intended spatial change, task consistency, answer validity, image dimensions, and file integrity.

Model-based curation. We further evaluate the candidate pool using Doubao Seed 2.1 Pro ByteDance Seed (2026), MiniMax M3 (Lai et al., 2026), and Qwen3-VL Plus (Bai et al., 2025a). For model m, let $\hat{y}^{\mathrm{O}}_{i,m}$ and $\hat{y}^{\mathrm{CF}}_{i,m}$ denote predictions on the original and counterfactual images. We remove samples satisfying

$$\mathcal{R} = \bigcap_{m=1}^{3} \left\{ i : \hat{y}^{\mathrm{O}}_{i,m} \neq y^{\mathrm{O}}_{i} \ \wedge \ \hat{y}^{\mathrm{CF}}_{i,m} = y^{\mathrm{CF}}_{i} \right\}. \tag{7}$$

This criterion removes samples for which all three models fail on the original image but succeed on the counterfactual image, which is inconsistent with the intended original-easier/counterfactual-harder setting. A total of 43 pairs (3.32% of the candidate pool) are removed, leaving 1,254 pairs for evaluation. This filtering serves only as a consistency check and does not assume that every retained counterfactual example is intrinsically more difficult than its original counterpart.

F.1.3 Construction Prompts

The construction prompt contains a general system prompt, a per-image user template, task-specific modules, few-shot examples, and successive expansion prompts. We provide a condensed, normalized rendering below to expose the common protocol.

Consolidated construction protocol. The protocol below summarizes the common requirements of the prompt family.

```
ROLE
Construct paired original/counterfactual aerial-image questions that
    test visible spatial evidence under a specified prior conflict.

INPUT
Source image; optional detection annotations; spatial task; bias
    reason; target/reference or target-group specification; intended
    original and counterfactual states; output ID; answer-position
    schedule; optional few-shot schema examples.

PROCEDURE
1. Inspect the source and annotations. Ground an identifiable target
    and any unchanged reference. Use a description that remains valid
    in both images. Reject an ambiguous target.
2. Preserve the original. Edit the supplied image rather than
    synthesizing a replacement scene. Intervene on the designated
    spatial factor of the target or target group. Preserve camera,
    crop, identity and unrelated content as far as the editing process
    permits.
3. Remove the old target footprint when relocating it. Avoid
    duplication, ghosting, unintended changes, and annotation overlays.
    Preserve lighting, support, perspective and occlusion unless the
    specified factor tests that property.
4. Inspect the realized edit. Determine the after-state from the
    pixels. Retry or reject if the required distinction is not visible
    or the target cannot be identified.
5. Write one shared four-option single-choice question. Use parallel,
    non-overlapping options and an explicit reference frame where
    required. Do not mention original/counterfactual status. Avoid
    unsupported claims about motion, intent, legality or metric depth.
6. Assign one answer per image. Require an answer change for newly
    constructed pairs. Apply the supplied option permutation to the
    options and both answer labels consistently.
7. Return the core record and available provenance/QC metadata. Report
    a rejection reason if construction fails; do not claim a
    nonexistent edited file.

CORE RECORD
```

```
id; bias_spatial; bias_reason; O_image_path; CF_image_path;
question; question_type="single-choice"; options={A,B,C,D};
O_answer; CF_answer; metadata.

PROVENANCE WHEN AVAILABLE
Source dataset/frame/annotation; target and reference grounding;
changed factor; original/counterfactual predicates; editor prompt;
editor identity; dimensions; checks; hashes; rejected attempts.
```

Task-specific modules. This module specifies the question's spatial predicate and guards against conflating adjacent tasks.

```
SPATIAL POSITION
Assign a target to an absolute image region or a named scene region.
    Avoid turning the question into a binary target/reference relation.

SPATIAL RELATIONSHIP
Ask for a target's relation to a fixed reference, such as image-plane
    direction, separation, contact, support or overlap.

SPATIAL ACTION
Ask for visibly depicted maneuver geometry or bodily action. A still
    image alone does not prove velocity, future intent or a temporal
    trajectory.

GEOMETRIC STRUCTURE
Classify a multi-object global configuration or a specified structural
    motif. Do not substitute a single object's position or orientation.

DEPTH AND SCALE
State the comparison rule: projected footprint or visibly supported
    depth ordering. Do not equate image size with metric distance.

VIEWPOINT AND ORIENTATION
Express the object's physical front or body axis in a declared
    image/scene frame. Preserve the camera when intervening on object
    orientation.
```

Prior-source modules. One module specifies the intended prior contrast. Its label is a construction hypothesis rather than an independently verified cognitive mechanism.

```
CULTURAL AND BEHAVIORAL NORMS
Contrast a common scene convention with a visibly different state. Ask
    about spatial evidence, not an unsupported legal or moral judgment.
    Do not invent jurisdiction-specific rules.

PHYSICAL-GEOMETRIC PRINCIPLES
Manipulate the designated support, contact, scale, perspective or
    orientation variable. Some historical examples deliberately create
    strong physical conflicts; later prompts prefer milder anomalies.

SEMANTIC CO-OCCURRENCE
Retain scene/category cues while making the tested object-layout,
    object-action or object-scale association less typical. The answer
    must follow the visible state rather than the usual association.

EGOCENTRIC-ALLOCENTRIC TRANSFORMATION
Require a mapping between object/scene coordinates and fixed image
    coordinates. Name the reference frame explicitly and ground its
    axes in visible evidence.
```

Per-image editing template. The abstract template is instantiated with the selected target, spatial states, and preservation constraints.

```
Edit the supplied image. Target: <stable target/group descriptor and
    source bounding box or annotation rows>.
Change only <designated spatial variable> from <observed state> to
    <requested state>.
Keep <fixed reference, camera, crop, identity, count and non-target
    context> unchanged. Reconstruct the vacated region and any required
    local contact/shadow details.
Do not add labels, boxes, arrows, text, duplicate targets or unrelated
    objects. Output one edited image for inspection.
```

Exampled editor instructions. Following shows several representative editor instructions for different kind of data generation.

Spatial position: `SPC_002`.

```
Move only the blue upper-left taxi diagonally downward and slightly
    right, without rotating it, until it clearly straddles both solid
    yellow centerlines. Reconstruct the old asphalt; keep the other
    taxi, road markings, cracks, camera, lighting, crop, and aspect
    ratio unchanged.
```

Spatial relationship: `SRP_001`.

```
Move only annotation row 85 onto the roof of fixed row 86; retain both
    colors, sizes, headings and visible silhouettes.
```

Spatial action: `SAC_003`.

```
Rotate only the green bus at [933,322,349,197] into an upward left-turn
    pose within the open junction, translating minimally to avoid all
    fixed cars, pedestrians, crossings, and the traffic island.
```

Geometric structure: `GSSC_005`.

```
Target only annotation rows 3,4,5,10. Reposition those four cars into a
    clear compact 2-by-2 rectangular valet-staging array, two rows of
    two with safe gaps, while preserving colors/types, perspective,
    shadows, camera and non-target content.
```

Depth and scale: `DSC_001`.

```
Enlarge only the lowest white vehicle in the central chain near
    [542,217,45,42] to about 170x125 pixels; preserve its center,
    heading, color, category, ground contact, the silver reference car,
    and all other content.
```

Viewpoint and orientation: `EAT_002`.

```
Use case: precise-object-edit. Image 1 is the original edit target.
    Rotate ONLY the large motion-blurred white panel van in the
    lower-left intersection at source bbox [259,403,268,213] by exactly
    180 degrees in yaw around its current image-plane center. Make the
    cab, windshield and front point unmistakably toward the
    UPPER-RIGHT, with the cargo rear toward the lower-left. Preserve
    the same plain white panel-van category, body proportions,
    projected area, center, scene depth, wheels-on-asphalt support,
    aerial perspective, natural motion-blur character, daylight and
```

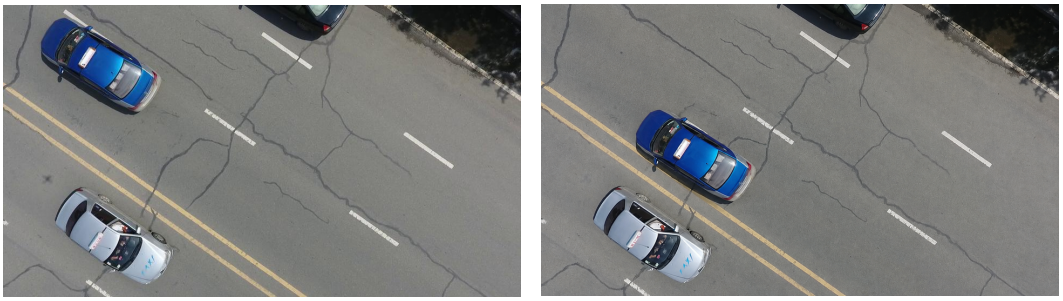
```
coherent shadow. Preserve every other car, pedestrian, signal,
crossing, barrier, tree, storefront, camera and crop. Reconstruct
minimal exposed asphalt. Do not translate, resize, deform,
duplicate, remove, add logos or readable text, rotate the whole
image or globally restyle. Output one photorealistic edited image
only.
```

F.2 Dataset Examples

This section presents several representative data samples from our spatial bias benchmark. Fig. 12–17 present six illustrative pairs, one for each spatial task and collectively covering all four bias-reason labels. The original and counterfactual images are displayed side by side for exposition; the shared question can be answered from either image individually. The displayed questions, option orders, and ground-truth labels are reproduced without rewriting from the final annotations.

The examples distinguish closely related spatial judgments. In Fig. 12, relocating a taxi changes its region membership, whereas Fig. 13 changes the support/contact relation between two cars. Fig.14 changes a bus's visible maneuver geometry. Fig. 15 rearranges a group of cars, testing global configuration rather than the position of a single object. Fig. 16 explicitly compares projected footprints, avoiding an unsupported inference from pixel size to metric depth. Finally, Fig. 17 requires expressing a van's physical forward axis in image coordinates. The physical-conflict examples intentionally include strong anomalies; the benchmark also contains less extreme, scene-plausible edits.

Original (O) **Counterfactual (CF)**



Question. Which road region contains the center of the blue taxi with the white roof sign?

A. The roadside shoulder outside the paved carriageways
B. The painted double-yellow centerline zone
C. The image-lower-left carriageway
D. The image-upper-right carriageway
Ground truth: O = **D**; CF = **B**.

Figure 12: **Spatial position / Cultural and behavioral norms** (`SPC_002`). The blue taxi is relocated from the upper-right carriageway onto the double-yellow centerline zone. The question asks for region membership rather than a legal judgment. The question, option order, and answer labels are reproduced from the final dataset.

Original (O) **Counterfactual (CF)**



Question. What is the physical spatial relationship between the bright red car in the lower-right diagonal row and the adjacent dark car immediately to its right?

A. The red car is following behind the dark car with a road gap
B. One of the two cars is not visible enough to determine the relation
C. The red car is resting directly on top of the dark car
D. Their vehicle bodies are separate in neighboring diagonal positions
Ground truth: O = **D**; CF = **C**.

Figure 13: **Spatial relationship / Physical-geometric principles** (`SRP_001`). The red car changes from a separate neighboring position to resting on the dark car. This deliberately strong support/contact intervention illustrates the physical-conflict subset. The question, option order, and answer labels are reproduced from the final dataset.

Original (O) **Counterfactual (CF)**



Question. Which current maneuver geometry best describes the bright green city bus in the intersection?

A. Crossing straight from image-left to image-right along the cross street
B. Making a U-turn within the horizontal cross street
C. Stopped parallel beside the central traffic island
D. Turning left from the cross street into the avenue toward the image top
Ground truth: O = **A**; CF = **D**.

Figure 14: **Spatial action / Cultural and behavioral norms** (`SAC_003`). The bus changes from a straight-across configuration to a left-turn configuration. The annotation concerns visible maneuver geometry; a single frame does not independently establish velocity or temporal motion. The question, option order, and answer labels are reproduced from the final dataset.

Original (O) **Counterfactual (CF)**

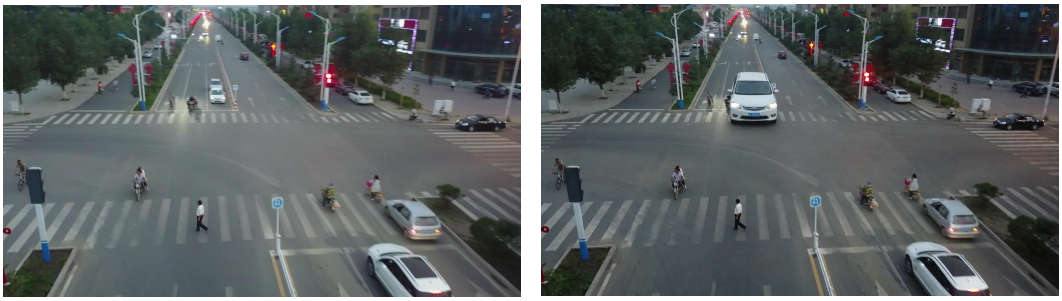


Question. In the storefront forecourt, what overall configuration is formed by the selected white, dark, white, and bronze cars?

A. An irregular scattered cluster
B. A two-by-two rectangular array
C. An alternating zigzag
D. One curb-parallel diagonal row
Ground truth: O = **D**; CF = **B**.

Figure 15: **Geometric structure / Semantic co-occurrence** (GSSC_005). The four selected cars change from a curb-parallel diagonal row to a two-by-two array. The target is a group configuration, with storefront context providing the hypothesized semantic prior. The question, option order, and answer labels are reproduced from the final dataset.

Original (O) **Counterfactual (CF)**



Question. Using projected image footprint as the decision rule, how does the lowest white vehicle in the central inbound chain, immediately above the far zebra crossing compare with the large silver hatchback crossing at the lower right?

A. At least one of the two objects is not visible enough to compare
B. Their projected footprints are approximately equal
C. The distant central white vehicle has a clearly larger projected footprint than the lower-right silver hatchback
D. The distant central white vehicle has a clearly smaller projected footprint than the lower-right silver hatchback
Ground truth: O = **D**; CF = **C**.

Figure 16: **Depth and scale / Physical-geometric principles** (DSC_001). The distant white vehicle is enlarged until its projected footprint exceeds that of the foreground silver hatchback. The question explicitly compares projected size rather than inferring metric depth. The question, option order, and answer labels are reproduced from the final dataset.

Original (O) **Counterfactual (CF)**



Question. When the white panel van's physical forward axis is expressed in the fixed aerial camera frame, which diagonal image direction does it point toward?

A. Toward the upper-right of the image
B. Toward the lower-right of the image
C. Toward the lower-left of the image
D. Toward the upper-left of the image
Ground truth: O = **C**; CF = **A**.

Figure 17: **Viewpoint and orientation / Egocentric-allocentric transformation** (`EAT_002`). The white panel van is rotated by approximately 180 degrees. Its physical forward axis must be mapped into the fixed image coordinate frame, changing from lower-left to upper-right. The question, option order, and answer labels are reproduced from the final dataset.