WHEN WORDS SPEAK LOUDER THAN IMAGES: TOWARDS UNDERSTANDING LANGUAGE BIAS IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite substantial progress across downstream applications, vision–language models (VLMs) remain susceptible to *language bias*, often prioritizing linguistic cues over visual evidence and consequently producing incorrect predictions. Prior studies have proposed various approaches to understanding and mitigating language bias in VLMs, yet their findings often conflict due to the difficulty of *tracing how language bias propagates within black-box VLMs*. Building on the word completion task, we trace how language bias propagates through VLM inference by (1) proposing a *diagnostic* framework that decomposes the inference process into four distinct yet interdependent stages to trace the propagation of language bias; and (2) examining how two key factors underlying language bias, i.e., *linguistic priors* and *cross-modal coverage*, evolve across these stages and ultimately give rise to incorrect predictions. The linguistic prior captures the strength of statistical bias induced by the language model component of a VLM and represents the origin of language bias, whereas cross-modal coverage measures the extent to which linguistic cues cover the visual content. By decomposing inference into four stages and characterizing the interplay between linguistic priors and cross-modal coverage across these stages, we propose a systematic framework for tracing the propagation of language bias throughout the inference process; and uncover the underlying mechanism of language bias by revealing the interplay between linguistic priors and cross-modal coverage.

1 INTRODUCTION

Vision–Language Models (VLMs) integrate a powerful Large Language Model (LLM) with a vision encoder to support tasks such as image captioning, visual question answering, and visually grounded dialogue. However, this architectural design can also introduce *language bias*, causing models to prioritize linguistic cues over visual evidence and potentially produce incorrect predictions (Liu et al., 2025b; Deng et al., 2025). For example, as illustrated in Figure 1, given the caption "*Some women are playing __ on the beach.*" a VLM may predict *volleyball* for the blank, consistent with a strong linguistic preference induced by the surrounding text. However, the image shows that the women are playing *soccer*, while the men are playing *volleyball*. Understanding why linguistic cues can override task-relevant visual evidence is therefore central to reliable multimodal understanding.

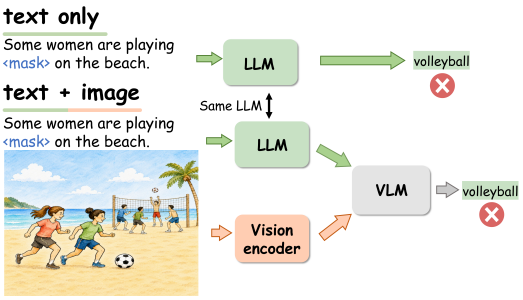


Figure 1: An example of language bias in VLMs: predictions from the underlying LLM can dominate the output, even when they conflict with the visual input.

There have been various studies in understanding and mitigating the language bias. One line of research *attributes biased predictions to specific architectural components*. For example, prior studies associate biased predictions with insufficient attention to visual information and show that amplify-

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.

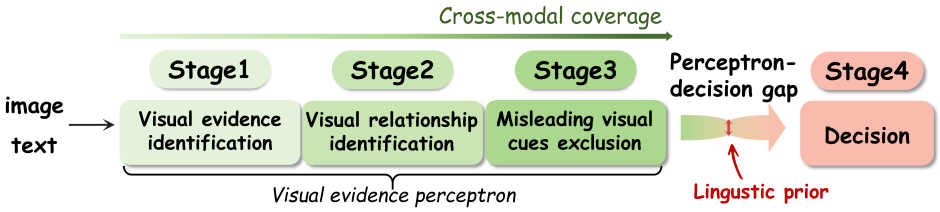


Figure 2: **Overview of the proposed four-stage diagnostic framework, illustrating how the interplay between cross-modal coverage and linguistic priors gives rise to language bias in VLMs.** The framework decomposes VLM inference into visual evidence identification (Stage 1), visual relationship identification (Stage 2), misleading visual cue exclusion (Stage 3), and final decision making (Stage 4). Across the first three stages, increasing cross-modal coverage enables the model to progressively extract and integrate visual evidence.

ing attention to image tokens can mitigate hallucinations (Liu et al., 2024; Yin et al., 2025). However, other studies find that manipulating visual attention does not necessarily eliminate VLMs' tendency to favor language bias (Yu et al., 2026; Ortu et al., 2026). Another line of research hypothesizes that *language bias stems from weak associations between visual and linguistic cues*, and seeks to mitigate such bias by strengthening these cross-modal associations, reporting improvements in hallucination mitigation for visual reasoning (Chen et al., 2023; Wu et al., 2026). But this approach fails in other tasks (Geigle et al., 2024; Wu et al., 2026). To understand the mechanisms underlying language bias, a prominent line of research leverages *counterfactual data* by constructing visual content that contradicts linguistic cues and examining how VLMs respond to such conflicting multimodal inputs (Liu et al., 2025b; Luo et al., 2025). However, this line of research relies heavily on the assumption that the multimodal data-generating function underlying VLMs can adequately account for counterfactual data, an assumption that remains unverified (Lin et al., 2023; Zhang et al., 2024; Jeong et al., 2026). More discussion about related works is available in Appendix A.1.

These seemingly conflicting findings, together with the difficulty of selecting appropriate analysis tools, stem from the ***black-box nature of VLMs***, making it challenging to trace how language bias propagates throughout the inference process (Feng et al., 2023). Toward understanding the mechanisms underlying language bias, this paper traces its propagation by **(1)** proposing a diagnostic framework that decompose the inference process into four distinct yet interdependent stages, providing a structured framework for tracing how language bias propagates throughout inference; and **(2)** focusing on two fundamental factors without relying on assumptions about specific architectural components. In addition, we focus on the *word completion* task. As illustrated in Figure 1, this task requires VLMs to predict a single word in response to a textual prompt, conditioned on the provided visual input. This is because the target word is typically a noun or adjective, which carries substantial lexical-semantic information (Baker, 2003; Baker & Croft, 2017), thereby facilitating our analysis of language bias.

As illustrated in Figure 2, this paper traces how language bias propagates throughout VLM inference by: **(1)** proposing ***a diagnostic framework*** that decompose the inference process into four distinct yet interdependent stages: visual evidence identification (Radford et al., 2021; Jia et al., 2021; Kamath et al., 2021; Li et al., 2023b), visual relationship identification (Yuksekgonul et al., 2023; Liu et al., 2023a; Hou et al., 2025; Chen et al., 2025), misleading visual cue exclusion (Mousi et al., 2026; Sun et al., 2026; Xie et al., 2025), and decision (Nooralahzadeh et al., 2026; Wang & Xu, 2026), thereby enabling the analysis of where language bias emerges during prediction. These stages serve as a functional diagnostic decomposition of the task rather than a claim that VLMs internally execute inference through the same sequential process; and **(2)** identifying two key factors underlying language bias, ***linguistic prior*** and ***cross-modal coverage***, and characterizing their interplay with our diagnostic framework. The linguistic prior captures the strength of statistical bias induced by the LLM component within the VLM, representing the *origin* of the language bias. Cross-modal coverage measures the extent to which linguistic cues describe the visual content in a model-agnostic manner, thereby providing a means to evaluate the *cross-modality interactions* between the two modalities. By focusing on these two factors, we avoid the challenge of making assumptions about which architectural components within black-box VLMs contribute to language bias, while our empirical observations demonstrate the effectiveness of these two factors in understanding such bias.

Building on this setting, our empirical experiments and statistical analyses elucidate the mechanisms underlying language bias through a diagnostic framework and the interplay between two key factors. Our main findings are:

Diagnostic Framework: Four-Stage Decomposition of the Inference Process. The four-stage decomposition provides a unified framework for tracing language bias and reconciling conflicting findings in prior studies. **(1) Beyond association:** successful inference requires not only visual–linguistic associations but also the identification of relevant visual relationships and the exclusion of misleading cues, explaining the limited effectiveness of association-based mitigation. **(2) Beyond attention:** the need to exclude misleading cues further shows that greater visual attention alone cannot prevent VLMs from relying on the linguistic prior.

Interplay between Linguistic Prior and Cross-Modal Coverage. Their interplay across the four stages reveals how linguistic priors diminish the effects of cross-modal coverage and ultimately lead to language bias. **(1) Cross-modal coverage facilitates progression:** higher cross-modal coverage is associated with progression to later inference stages. **(2) Linguistic priors create an perceptron–decision gap:** they can diminish the benefits of cross-modal coverage before the decision stage, allowing correctly recognized visual evidence to be overridden in the final decision.

Organization. § 2 introduces the preliminaries for uncovering the mechanisms underlying language bias, including the diagnostic framework, linguistic priors, and cross-modal coverage. § 3 presents a mechanistic analysis of language bias. § 4 discusses the key findings and explores potential approaches to mitigating language bias. Finally, § 5 concludes the paper.

2 PRELIMINARY

In this section, we establish the preliminaries for analyzing the mechanisms underlying language bias. Specifically, we first introduce the dataset and task formulation, then present a diagnostic framework that decomposes the inference process into four stages (§ 2.1), and finally formalize the notions of linguistic prior and cross-modal coverage (§ 2.3 and § 2.4).

We use the dataset developed by Ma et al. (2023), which, as illustrated in Figure 1, requires VLMs to perform a word-completion task based on multimodal inputs.

Notations. Let x_v denote the image, x_l the text containing a blank, and w^{gold} the ground-truth completion. Given the multimodal input (x_v, x_l), a VLM f parameterized by θ produces a completion: $\hat{w}^{\text{vl}} = f_\theta(x_v, x_l)$. Given the text alone, f produces a text-only prediction: $\hat{w}^{\text{l}} = f_\theta(x_l)$. We use $P^{\text{l}}_\theta(w \mid x_l)$ to denote the probability assigned to a candidate completion w under the text-only setting. We further use p_{ling} to denote the linguistic-prior score and p_{cov} to denote cross-modal coverage.

2.1 A FOUR-STAGE DIAGNOSTIC FRAMEWORK

Examining how language bias propagates during VLM inference is challenging due to the black-box nature of VLMs. Motivated by prior work on the mechanistic analysis of VLMs, we introduce a diagnostic framework that decomposes the inference process into four stages. We provide empirical evidence showing that our framework can effectively distinguish multimodal examples exhibiting language bias from those without such bias. Figure 3 illustrates what each stage measures in our four-stage diagnostic framework, as detailed below:

S1: Visual evidence identification. S1 measures to what extent VLMs can adequately identify the visual evidence of the ground-truth noun completion, e.g., soccer. Such cross-modal association is widely regarded as a fundamental prerequisite for effective visual-language grounding (Ma et al., 2023; Vong et al., 2024).

S2: Visual relationship identification. S2 measures the extent to which VLMs can correctly identify the visual relationships associated with the ground-truth answer, e.g., the relationships between women and soccer and between men and volleyball. Prior studies have demonstrated that accurately capturing such visual relationships is crucial for effective vision-language understanding and reasoning (Hou et al., 2025; Liu et al., 2023a).

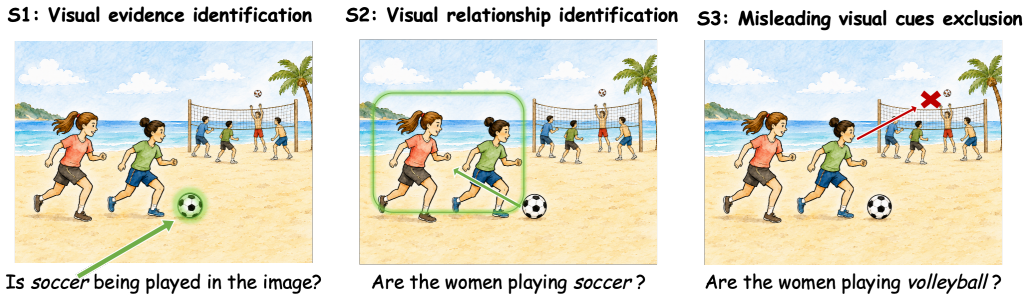


Figure 3: **Visualization of what each stage (the first three stages) measures in our diagnostic framework.** *S1: Visual evidence identification.* Assesses whether VLMs can identify the visual evidence corresponding to the ground-truth completion. *S2: Visual relationship identification.* Assesses whether VLMs can correctly associate the ground-truth visual evidence with its relevant visual cues. *S3: Misleading visual cue exclusion.* Assesses whether VLMs can reject misleading visual cues. **Notably**, in our diagnostic framework, *S4: Decision* corresponds directly to the VLM's prediction process and is therefore not included in the figure.

S3: Misleading visual cue exclusion. Even when VLMs correctly identify the relevant visual relationships, their predictions can still be misled by distracting or conflicting visual cues. In particular, correctly recognizing a true visual relationship does not necessarily imply that the model can reject a plausible but false alternative (Sun et al., 2026; Mousi et al., 2026). Building on S2, S3 evaluates whether VLMs can identify and resist misleading visual cues that may interfere with an otherwise correct visual relationship. For example, in Figure 1, volleyball serves as the misleading visual cue.

S4: Decision. We further introduce a dedicated *decision* stage to characterize how VLMs translate multimodal evidence into a final prediction. This distinction is motivated by mechanistic studies showing that constructing task-relevant internal representations and effectively leveraging those representations for prediction are distinct processes (Geva et al., 2023; Neo et al., 2025). In other words, a model may form the correct internal representation without successfully translating it into the correct prediction. In our implementation, the decision stage corresponds directly to the VLM's final prediction of the target word completion given the multimodal input.

As illustrated in Figure 2, we regard the first three stages associated with visual cues as the ***visual evidence perception*** process, which characterizes how VLMs perceive and extract evidence from visual inputs. We then focus on the decision stage, investigating how the perceived visual evidence can be leveraged to prevent the propagation of language bias.

2.2 EVALUATION OF VLM PERFORMANCE AT EACH STAGE

Following prior work on prompt-based probing of multimodal models (Salin et al., 2022; Cao et al., 2022; Zhao et al., 2024; Zhou et al., 2025; Hou et al., 2025), we use natural language probes to evaluate VLM performance at each stage of our diagnostic framework.

S1: Visual Evidence Identification	**S3: Misleading Visual Cue Exclusion**
Is *soccer* being played in the image?	Q1: Are the women playing *volleyball*?
S2: Visual Relationship Identification	Q2: Are the men playing *soccer*?
Q1: Are the women playing *soccer*?	**S4: Decision**
Q2: Are the men playing *volleyball*?	The original multimodal input.

Example 1: Stage-wise probing questions for the example in Figure 1.

Probing questions. Example 1 illustrates the probing questions for the example shown in Figure 1. The probing questions are designed to be straightforward and directly derived from the definition of each stage. For S1 (visual evidence identification), we directly probe whether the ground-truth word is visually present in the image. Both S2 and S3 contain two probing questions. For S2 (visual relationship identification), the questions assess whether the VLM can identify the correct visual relationship associated with the ground-truth completion. In contrast, S3 probes potentially misleading visual relationships, for which the expected answer is *No*, allowing us to examine whether visually salient but irrelevant evidence biases the model's reasoning. Finally, for S4, we directly

Model	Biased (%)			No-bias (%, Δ)		
	S1	S2	S3	S1	S2	S3
Gemma 3 4B	92.9	87.5	16.1	98.7 (+5.8)	94.4 (+6.9)	59.8 (+43.7)
Qwen2.5-VL 7B	64.1	74.4	25.6	78.2 (+14.1)	89.4 (+15.0)	57.6 (+32.0)
OneVision 1.5 4B	86.4	88.6	25.0	93.9 (+7.5)	95.8 (+7.2)	52.7 (+27.7)
OneVision 1.5 8B	83.8	83.8	29.7	95.4 (+11.6)	97.5 (+13.7)	61.1 (+31.4)

Table 1: **VLM performance across Stages S1–S3 under biased and unbiased conditions.** Values in parentheses denote the performance gap ΔS_k, computed as the no-bias minus biased performance at stage S_k. S1, S2, and S3 denote Visual Evidence Identification, Visual Relationship Identification, and Misleading Visual Cue Exclusion, respectively.

evaluate the VLM's output given the complete multimodal input. The complete prompt templates and additional examples are provided in Appendix A.2.

Validation of the diagnostic framework. To further validate the four-stage design of our diagnostic framework, Table 1 compares stage-wise performance across four VLMs for cases with and without language bias. In both conditions, the text-only model produces an incorrect completion. In *biased* cases, the VLM retains this linguistically driven incorrect prediction even after receiving the image, indicating the presence of language bias. In *no-bias* cases, the VLM instead uses the visual evidence to correct the initial prediction, indicating the absence of language bias. This provides a rigorous and stringent criterion for determining whether a case exhibits language bias.

Across all four VLMs, ***the no-bias cases consistently achieve higher performance than the biased cases at each of the three stages, providing empirical support for the effectiveness of our diagnostic framework***. The average performance gaps are 9.8, 10.7, and 33.7 percentage points at S1, S2, and S3, respectively. This consistent stage-wise separation demonstrates that the framework effectively distinguishes cases in which VLMs successfully use visual evidence from those in which linguistic priors lead to biased predictions. The particularly large gap at S3 further indicates that misleading visual cue exclusion is a critical stage in the emergence of language bias.

2.3 Linguistic Prior

Linguistic prior has been widely recognized as ***a major source of language bias*** in VLMs (Goyal et al., 2017; Wu et al., 2023; Lin et al., 2024a; Lee et al., 2025; Deng et al., 2025). Here, we use linguistic prior to measure the strength of statistical bias induced by the language model component of a VLM. We define the linguistic-prior score as:

$$p_{\text{ling}} = \frac{P_\theta^{\text{l}}(\hat{w}^{\text{l}} \mid x_l)}{P_\theta^{\text{l}}(\hat{w}^{\text{l}} \mid x_l) + P_\theta^{\text{l}}(w^{\text{gold}} \mid x_l)} \tag{1}$$

The score p_{ling} quantifies the text-only model's relative preference for its prediction $\hat{w}^{\text{l}}$ over the ground-truth completion w^{gold}, with larger values indicating a stronger preference for $\hat{w}^{\text{l}}$.

2.4 Cross-modal Coverage

Cross-modal coverage measures how much of the visual cues relevant to the ground-truth word completion is already expressed in the textual input. However, identifying which visual cues are relevant to the ground-truth completion is non-trivial. An image typically contains more visual information than can be captured by a textual description (Elliott & de Vries, 2015; Tavakoli et al., 2017; Ilinykh et al., 2018; Kreiss et al., 2022; Chan et al., 2023), textual descriptions are selective, verbalizing only a subset of the visual content rather than describing everything in the image. We therefore ***employ human annotators to identify visual cues relevant to the ground-truth completion*** and determine which of these cues are already expressed in the textual input.

To estimate cross-modal coverage, we develop a three-step annotation pipeline, as illustrated in Figure 4. First, we prompt off-the-shelf LLMs to verbalize the visual cues present in the visual input, providing candidate cues for subsequent annotation. Second, given the visual input, textual input, and ground-truth word completion, human annotators identify the verbalized visual cues relevant to the ground-truth completion and add any relevant cues missed during the verbalization step. We

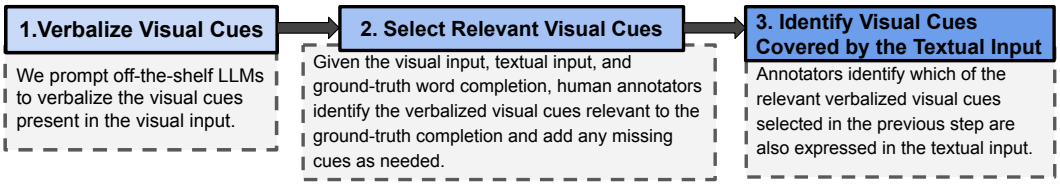


Figure 4: **Overview of our annotation pipeline for cross-modal coverage.** We first prompt off-the-shelf LLMs to verbalize the visual cues present in the visual input. Human annotators then identify the cues relevant to the ground-truth word completion and add any missing cues. Finally, annotators determine which of these relevant visual cues are already expressed in the textual input.

denote the number of relevant visual cues by N_V^r. Finally, annotators determine which of these relevant visual cues are already expressed in the textual input, and we denote the number of such covered cues by N_L^r. This procedure allows us to distinguish between relevant visual cues that are already covered by the textual input and those that remain available only in the visual input.

Cross-modal coverage is then defined as: $p_{\text{cov}} = N_L^r / N_V^r$**.** Thus, p_{cov} measures the proportion of visual cues relevant to the ground-truth word completion that are also expressed in the textual input. Further annotation details are provided in Appendix A.3.

3 Mechanistic Analysis

In § 2, we introduce the methodological preliminaries for analyzing the mechanisms underlying language bias. In this section, we first assess linguistic prior and cross-modal coverage as informative variables by testing their associations with final prediction correctness. Using our diagnostic framework, we then examine their roles in visual evidence perception and the final decision. Our findings are: (1) cross-modal coverage enhances visual evidence perception; and (2) linguistic priors can override the benefits of cross-modal coverage.

Experimental setup and backbone models. We evaluate five instruction-tuned VLMs from three model families: Gemma 3 (4B and 12B) (Gemma Team, 2025), Qwen2.5-VL 7B (Bai et al., 2025), and LLaVA-OneVision 1.5 (4B and 8B) (An et al., 2025). All models are evaluated on the same set using human-reviewed correctness that accepts semantically valid completions consistent with both inputs.

Table 2: Associations of cross-modal coverage and linguistic prior with completion correctness. β_C and β_P denote the coefficients of cross-modal coverage and linguistic prior, respectively. For Qwen2.5-VL 7B, we additionally report results in low- and high-prior regimes. $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

Model	Items	β_C	β_P
Gemma 3 4B	2,839	1.409^{***}	-1.740^{***}
Gemma 3 12B	2,839	2.209^{***}	-1.910^{***}
Qwen2.5-VL 7B	2,839	0.080	-1.609^{**}
OneVision 1.5 4B	2,839	1.165^{***}	-3.012^{***}
OneVision 1.5 8B	2,839	1.082^{***}	-3.713^{***}

3.1 Cross-Modal Coverage Enhances Visual Evidence Perception

Prior studies have linked visually ungrounded predictions to excessive reliance on language priors (Favero et al., 2024; Lin et al., 2024a) and insufficient use or grounding of visual evidence (Favero et al., 2024; Dai et al., 2023). Motivated by these two lines of work, we focus on two directly measurable variables for language bias: *linguistic prior* and *cross-modal coverage*.

To characterize the role of cross-modal coverage within our diagnostic framework, we ***examine how greater cross-modal coverage enhances visual evidence perception*** and enables VLMs to progress

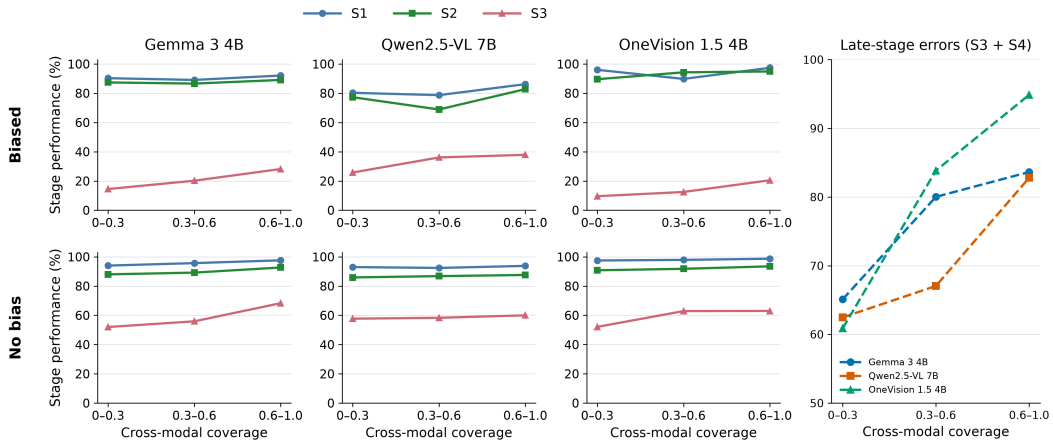


Figure 5: Diagnostic-stage performance and error distribution across cross-modal coverage intervals for one representative model from each model family. Left: S1–S3 performance on noun-target items for biased (top) and no-bias (bottom) cases across Gemma 3 4B, Qwen2.5-VL 7B, and OneVision 1.5 4B. Right: the proportion of noun-target completion errors assigned to S3 or S4 across the same coverage intervals. Results for Gemma 3 12B and OneVision 1.5 8B are provided in Appendix A.4.

through the first three stages. To this end, we conduct statistical analyses to examine (1) whether greater cross-modal coverage improves VLM performance across the first three stages in both biased and no-bias cases and (2) whether it facilitates VLM progression from S1 to S3.

We first examine whether cross-modal coverage and linguistic prior are systematically associated with final prediction correctness. A logistic regression jointly considering the two factors shows opposing associations with correctness: greater cross-modal coverage is positively associated with correctness for four of the five models, whereas stronger linguistic prior is negatively associated with correctness across all five models (Table 2). Together, these results provide statistical support for cross-modal coverage and linguistic prior as two informative factors associated with final prediction behavior.

We next examine how this relationship is reflected within our diagnostic framework. For this stage-wise analysis, we only include items whose ground-truth completion is a noun (NOUN). Noun targets allow concrete evaluation of the relevant visual concept, its relationship to the queried entity or action, and the validity of a candidate completion. Figure 5 shows that as cross-modal coverage increases, performance across the diagnostic stages generally improves, with the clearest change appearing at S3. This pattern is observed in both biased and no-bias cases, indicating that greater cross-modal coverage is associated with stronger visual evidence perception. For clarity, Figure 5 presents one representative model from each model family; results for Gemma 3 12B and OneVision 1.5 8B are provided in Appendix A.4 and show the same qualitative pattern.

At the same time, among incorrect noun-target completions, the share of errors assigned to S3 or S4 increases with cross-modal coverage. Even at higher coverage levels, difficulties with misleading visual cue exclusion and the final decision remain prominent among the errors that persist. S4 cases are particularly informative: on some items, the model answers all S1–S3 probes correctly yet completes the original caption incorrectly. We examine these cases in the next section.

3.2 The Linguistic Prior Overwrites Visual Evidence Perceptron

In the previous section, we showed that higher cross-modal coverage is associated with better performance on S1–S3. We now examine the role of linguistic priors in the perceptron–decision gap, which describes cases in which a VLM answers the S1–S3 probes correctly but still makes an incorrect final prediction. Specifically, we ask whether ***strong linguistic priors can still bias the final prediction when the model has answered the visual-evidence probes correctly***.

Table 3: Final completion accuracy among cases with correct S1–S3 judgments for $p_{\text{ling}} < 0.99$ and $p_{\text{ling}} \geq 0.99$. Δ denotes the change in accuracy from the former to the latter, in percentage points.

Model	$p_{\text{ling}} < 0.99$	$p_{\text{ling}} \geq 0.99$	Δ
Gemma 3 4B	98.3% ($n = 58$)	95.6% ($n = 135$)	-2.7
Gemma 3 12B	94.1% ($n = 185$)	92.3% ($n = 750$)	-1.8
Qwen2.5-VL 7B	95.7% ($n = 69$)	91.8% ($n = 146$)	-3.9
OneVision 1.5 4B	97.7% ($n = 43$)	89.9% ($n = 99$)	-7.8
OneVision 1.5 8B	92.6% ($n = 54$)	90.5% ($n = 116$)	-2.1

Before examining the counterfactual pairs, we first test whether strong linguistic priors remain associated with final-decision failures after S1–S3 have been successfully completed. As shown in Table 3, final accuracy is consistently lower in the very-high-prior regime across all five models, even when all S1–S3 judgments are correct, motivating a closer examination of the role of linguistic prior in the final decision.

To examine the effects of linguistic priors, we (1) construct adversarial textual inputs that preserve the semantics of the original inputs but induce different linguistic priors; and (2) apply these inputs to the VLMs to analyze how changes in linguistic priors affect their final decisions. Constructing such examples is challenging because the linguistic prior must be altered without changing cross-modal coverage. To address this challenge, ***we restrict modifications to adjectives, prepositional phrasing and verbs***, prompt off-the-shelf LLMs to generate candidate examples, and retain only those that induce a weaker linguistic prior while preserving cross-modal coverage and successful S1–S3 judgments. For models with insufficient perceptron–decision gap cases, multiple captions associated with the same image are used, with each caption reformulated and evaluated independently. One such pair is shown below:

Original: "A girl is smelling a mushroom that a woman is holding up to her."
Adversarial: "A girl is smelling a mushroom that a woman is holding in front of her."

Both textual inputs express the same visual cues and therefore have the same cross-modal coverage, but they can induce different linguistic priors. We construct 100 reformulation pairs for each model, resulting in 500 pairs across the five evaluated models. Across these controlled examples, weakening the linguistic prior can change the final completion even though cross-modal coverage and S1–S3 judgments remain unchanged. Additional prior-counterfactual examples, including the corresponding changes in linguistic prior and final completion, are provided in Appendix A.5.

Together, the prior-stratified observation and the controlled reformulations provide complementary evidence that linguistic prior is involved in the perceptron–decision gap. Even after successful visual evidence perception, very strong linguistic priors are associated with lower final completion accuracy, and weakening the linguistic prior can change the final decision while cross-modal coverage and S1–S3 judgments remain unchanged. Our construction is intended to demonstrate that such cases can occur rather than to estimate their frequency in the full evaluation set. Because reformulation also changes the surface form of the textual input, these examples do not establish linguistic prior as the sole cause of the perceptron–decision gap.

4 Discussion

Language bias in current VLMs may require more challenging settings. Our results suggest that language bias is increasingly difficult to observe in stronger VLMs under relatively straightforward visual–linguistic conflicts. For example, larger models can often use available visual evidence to override misleading linguistic preferences, whereas smaller models exhibit more persistent language-biased behavior. This observation suggests that future evaluations should move beyond simple modality conflicts and consider more challenging cases, where visual evidence is incomplete, ambiguous, distributed, or requires compositional reasoning. Such settings may better reveal when strong linguistic priors continue to interfere with visual evidence even in capable VLMs.

Grounding based on statistical associations may itself introduce bias. A fundamental challenge is that current vision–language grounding is largely built upon statistical associations learned from

large-scale image–text datasets. These associations provide powerful priors that enable VLMs to recognize common concepts and generate fluent responses, but they can also encourage models to rely on correlations that are not aligned with the specific visual evidence in a given instance. Therefore, language bias should not be viewed only as an undesirable error introduced after training; it may partially originate from the same statistical mechanisms that enable effective multimodal learning.

The dual role of linguistic priors in VLMs. Our findings reveal a fundamental trade-off in mitigating language bias. The linguistic prior that contributes to biased predictions under visual conflicts is also a major source of the knowledge and generalization ability provided by large language models. Therefore, reducing language bias by suppressing linguistic influence may also limit the ability of VLMs to leverage the broad knowledge encoded in their language models. This trade-off suggests that language bias cannot be viewed as an isolated failure mode independent of the capabilities that make VLMs effective.

Revisiting vision–language grounding for eliminating language bias. Our findings suggest that reducing language bias may require going beyond statistical associations as the basis of vision–language grounding. Current VLMs largely acquire semantic alignment from large-scale image–text co-occurrences, which provides strong generalization but can also entangle useful linguistic knowledge with instance-level biases. A more fundamental direction may be to construct semantic spaces that better reflect linguistic structures and cognitive principles, allowing visual and linguistic information to be organized according to their functional roles rather than their statistical correlations alone. Such grounding mechanisms could potentially reduce language bias while preserving, or even improving, the knowledge utilization ability of VLMs.

5 Conclusion

Final answers alone leave the source of language bias in VLMs difficult to locate. We study this question with a four-stage diagnostic framework for a word-completion task, using cross-modal coverage and linguistic prior to characterize biased behavior. Higher cross-modal coverage is associated with better performance across S1–S3, yet correct judgments at these stages do not guarantee a correct final completion; controlled caption reformulations further show that reducing linguistic prior can alter the final decision while visual evidence judgments remain unchanged. By separating visual evidence perception from final decisions, our framework provides a finer-grained view of where language bias persists in VLM inference. More broadly, simply suppressing linguistic priors may reduce access to the rich linguistic knowledge inherited from large language models. A more fundamental solution may require developing semantic spaces for vision–language grounding that better reflect linguistic structures and plausible principles of human cognition, enabling visual evidence and linguistic knowledge to be integrated rather than competing with each other.

AI use statement

We used generative AI tools to generate candidate textual reformulations for the prior-counterfactual experiments, as described in Section 3.2. We also used generative AI tools for research brainstorming, language editing, and limited assistance with code drafting and debugging. All AI-generated experimental inputs were filtered according to the predefined experimental criteria and manually reviewed before use. AI-assisted code and analyses were checked by the authors, and all reported results and scientific claims were independently verified. Generative AI was not used to make final human annotation or correctness decisions. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work studies language bias in vision-language models using an existing benchmark and does not involve sensitive personal data or deployment on human subjects. Our analysis is intended to improve understanding of how linguistic priors affect multimodal predictions rather than to make claims about human behavior or social groups. Human annotations are used only for evaluating task-relevant visual and textual information and model correctness. We do not release personally identifiable information or other sensitive content.

Reproducibility Statement

We provide detailed definitions of linguistic prior, cross-modal coverage, and the diagnostic framework in Section 2. The annotation procedure, additional stage-wise results, and prior-counterfactual examples are provided in the appendix A. We also report the evaluated models, experimental settings, and statistical analyses used throughout the experiments.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. Technical report, Alibaba Group, 2025. arXiv:2502.13923.

Mark Baker and William Croft. Lexical categories: Legacy, lacuna, and opportunity for functionalists and formalists. *Annual Review of Linguistics*, 3:179–197, 2017. doi: 10.1146/annurev-linguistics-011516-034134.

Mark C. Baker. *Lexical Categories: Verbs, Nouns and Adjectives*. Cambridge University Press, Cambridge, 2003. doi: 10.1017/CBO9780511615047.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding Information Storage and Transfer in Multi-Modal Large Language Models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0dfe31d6e703e138d46a7d2fced38b7c-Abstract-Conference.html.

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5796–5808. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.398. URL https://aclanthology.org/2022.acl-long.398/.

David M Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. Ic3: Image captioning by committee consensus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8975–9003, 2023.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. URL https://arxiv.org/abs/2306.15195.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas. In *Proceedings of the International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/chen25cr.html.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating. In *Proceedings of the 43rd International Conference on Machine Learning (ICML), Spotlight*, 2026. URL https://arxiv.org/abs/2605.24024. arXiv:2605.24024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.),

International Conference on Learning Representations, volume 2024, pp. 54158–54183, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edc36117f795ca52a0cbf6a7b3882859-Paper-Conference.pdf.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2136–2148, 2023.

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

Desmond Elliott and Arjen de Vries. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 42–52, 2015.

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

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11737–11762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.656. URL https://aclanthology.org/2023.acl-long.656/.

Gregor Geigle, Radu Timofte, and Goran Glavaš. Does object grounding really reduce hallucination of large vision-language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2728–2742. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.159. URL https://aclanthology.org/2024.emnlp-main.159/.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 66510–66547, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a6805b5564bd8d813a81c4b5a97e5ca6-Paper-Conference.pdf.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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, 2024. URL https://arxiv.org/abs/2310.14566.

Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23910–23959. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hou25c.html.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

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, 2024b. URL https://arxiv.org/abs/2311.17911.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. The task matters: Comparing image captioning and task-based dialogical image description. In *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 397–402, 2018.

Jongoh Jeong, Hoyong Kwon, Minseok Kim, and Kuk-Jin Yoon. Multimodal distribution matching for vision-language dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23072–23082, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR – modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Kamath_MDETR_-_Modulated_Detection_for_End-to-End_Multi-Modal_Understanding_ICCV_2021_paper.html.

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. Concadia: Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4667–4684, 2022.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.

In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual evidence prompting mitigates hallucinations in large vision-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.205. URL https://aclanthology.org/2025.acl-long.205/.

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. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Victoria Lin, Louis-Philippe Morency, Dimitrios Dimitriadis, and Srinagesh Sharma. Counterfactual augmentation for multimodal learning under presentation bias. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 592–606. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.43.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29914–29934. PMLR, 2024a.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29914–29934. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/lin24c.html.

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/.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291/.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b4008025c2182bfe16fcc8566ee14d64-Paper-Conference.pdf.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2407.21771.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight: Exploring the vision-knowledge conflicts in multimodal LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17825–17846, 2025b. doi: 10.18653/v1/2025.acl-long.872. URL https://aclanthology.org/2025.acl-long.872/.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156. PMLR, 2025. URL https://proceedings.mlr.press/v267/luo25b.html.

Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 524–544, 2023. doi: 10.18653/v1/2023.acl-long.31. URL https://aclanthology.org/2023.acl-long.31/.

Basel Mousi, Fahim Dalvi, Shammur Absar Chowdhury, Firoj Alam, and Nadir Durrani. Once correct, still wrong: Counterfactual hallucination in multilingual vision-language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4763–4788, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.234. URL https://aclanthology.org/2026.findings-acl.234/.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/900fb3439e4968df79a6f2bfedec49cd-Abstract-Conference.html.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026. URL https://arxiv.org/abs/2604.09364.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14109–14130, 2026. doi: 10.18653/v1/2026.acl-long.642. URL https://aclanthology.org/2026.acl-long.642/.

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

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 2022. doi: 10.1609/aaai.v36i10.21375.

Yizheng Sun, Mochuan Zhan, Yanan Ma, Jia Tong See, Yifan Wang, Ziyi Wang, Hao Li, Yang Cui, Wenhao Cai, Jingyu Sun, Chenghua Lin, Riza Batista-Navarro, and Jingyuan Sun. Are Reasoning Vision-Language Models Robust to Semantic Visual Distractions? *arXiv preprint arXiv:2606.08894*, 2026. URL https://arxiv.org/abs/2606.08894.

Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2487–2496, 2017.

Wai Keen Vong, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zeyu Wang and Xinming Xu. Knowing isn't always saying: When do spatial encodings reach answers in vision-language models? *arXiv preprint arXiv:2608.22916*, 2026. URL https://arxiv.org/abs/2608.22916.

Chenwei Wu, Li Erran Li, Stefano Ermon, Patrick Haffner, Rong Ge, and Zaiwei Zhang. The role of linguistic priors in measuring compositional generalization of vision-language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pp. 118–126. PMLR, 2023.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. PostAlign: Multimodal grounding as a corrective lens for MLLMs. In *International Conference on Learning Representations*, volume 2026, pp. 78875–78894, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7f52f6b8f107931127eefe15429ee278-Abstract-Conference.html.

Ming-Kun Xie, Jia-Hao Xiao, Gang Niu, Lei Feng, Zhiqiang Kou, Min-Ling Zhang, and Masashi Sugiyama. What Makes "Good" Distractors for Object Hallucination Evaluation in Large Vision-Language Models? *arXiv preprint arXiv:2508.06530*, 2025. URL https://arxiv.org/abs/2508.06530.

Hao Yin, Guangzong Si, and Zilei Wang. ClearSight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.13107.

Liu Yu, Can Chen, Ping Kuang, Zhikun Feng, Fan Zhou, and Gillian Dobbie. Dismantling pathological shortcuts: A causal framework for faithful LVLM decoding. *arXiv preprint arXiv:2606.27596*, 2026. URL https://arxiv.org/abs/2606.27596. Accepted at ICML 2026.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21853–21862, 2024.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13186–13214. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.771. URL https://aclanthology.org/2024.findings-emnlp.771/.

Kankan Zhou, Eason Lai, Kyriakos Mouratidis, and Jing Jiang. FOCUS: Evaluating pre-trained vision-language models on underspecification reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27565–27584. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1337. URL https://aclanthology.org/2025.acl-long.1337/.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fc625e831361cfcc82cb74224fdc66cb-Paper-Conference.pdf.

A Appendix

A.1 Related works

Language Bias in Vision–Language Models. Behavioral studies investigate how linguistic information influences VLM predictions when it conflicts with visual evidence. Explicit textual interference can lead models to follow misleading descriptions despite contradictory images, with this preference affected by text relevance, token order, and model scale (Deng et al., 2025). Conflicts can also arise without explicit misleading text. When images contradict commonsense knowledge, models sometimes answer according to their parametric knowledge rather than the depicted situation (Liu et al., 2025b). Related work has also examined linguistic shortcuts in image–text matching. Caption likelihood can inflate retrieval performance even without visual evidence (Lin et al., 2024b), while compositional evaluations reveal failures in attribute association, relational understanding, and word-order sensitivity that standard retrieval evaluations can overlook (Yuksekgonul et al., 2023). Dedicated benchmarks further examine these behaviors through capability controls and systematic changes to the inputs. VLind-Bench evaluates language dependence only after checking commonsense knowledge, visual perception, and commonsense bias, thereby reducing the influence of these confounding factors (Lee et al., 2025). Counterfactual and out-of-distribution evaluations contrast cases that can be answered from linguistic priors with cases that require visual discrimination, revealing substantial weaknesses when familiar linguistic associations no longer hold (Liu et al., 2025b; Luo et al., 2025). Controlled image–question groups further expose language hallucination, visual illusion, and inconsistencies across related responses (Guan et al., 2024). Together, these evaluations characterize not only whether predictions are incorrect, but also whether they follow linguistic cues, respond to changes in visual evidence, and remain consistent across related questions.

Mitigation methods address language bias and associated hallucinations through training, decoding, additional visual evidence, and output revision. PostAlign combines visual grounding with textual rationales and rejection of nonexistent objects to improve fine-grained visual understanding and reduce hallucinations (Wu et al., 2026). During generation, visual contrastive decoding contrasts predictions under original and distorted images to reduce reliance on statistical biases and unimodal priors (Leng et al., 2024). Visual evidence can also be made explicit through specialist-model outputs (Li et al., 2025) or image descriptions that guide subsequent decoding (Ghosh et al., 2025). At the output level, LURE revises generated descriptions using signals related to object co-occurrence, uncertainty, and position in the generated text (Zhou et al., 2024). However, mitigation gains depend on the evaluation setting: adding object-grounding objectives has little to no effect on hallucination in open-ended caption generation in the experiments of Geigle et al. (2024). We complement these evaluations by diagnosing language-biased behavior at the level of functional judgments, and then examine how these diagnostic patterns vary with linguistic prior and cross-modal coverage.

Mechanistic Analysis of Language and Vision–Language Models. Mechanistic studies examine how internal representations and computational components contribute to model predictions.

Information-flow analyses of factual recall distinguish the enrichment of subject representations from the extraction of queried attributes (Geva et al., 2023). Circuit analyses identify interacting attention heads and connections that support specific behaviors through both manual investigation and automated discovery (Wang et al., 2023; Conmy et al., 2023). The reliability of such explanations is itself an active area of study: RAVEL evaluates the disentanglement of attributes in distributed representations (Huang et al., 2024a), while activation-patching studies show that localization can depend on the evaluation metric and corruption method (Zhang & Nanda, 2024).

Other work directly intervenes on internal computation. Inference-Time Intervention steers activations toward directions associated with truthful answers (Li et al., 2023a), while DoLa exploits differences between layer-wise predictions to improve factual generation (Chuang et al., 2024). Related analyses also examine discrepancies between internal truth-related representations and generated responses (Liu et al., 2023b).

In multimodal models, causal tracing has been used to identify components involved in visual information transfer (Basu et al., 2024), while targeted attention-head interventions can shift predictions toward visual evidence or parametric knowledge (Ortu et al., 2026). Corresponding methods strengthen visual attention (Liu et al., 2024; Yin et al., 2025), modify attention patterns associated

with hallucination (Huang et al., 2024b), steer multimodal representations (Liu et al., 2025a), or separate visual and textual routes within attention heads (Cheng et al., 2026). Recent studies further show that visual information may remain encoded even when the final prediction contradicts it, and that its influence on the answer depends on where and how that information is used (Nooralahzadeh et al., 2026; Wang & Xu, 2026).

A.2 Diagnostic Prompt Templates and Examples

The diagnostic questions used in our experiments are instantiated from shared prompt templates. For each item, annotators identify the ground-truth concept, the competing concept, and the visual entities and relationships associated with them. The surface wording is adapted to the semantic type of each item, while the functional judgment evaluated at each stage remains fixed.

Stage	Shared Prompt Template
S1: Visual Evidence Identification	Is [GROUND-TRUTH CONCEPT] visually present in the image? For activity concepts: "Is [GROUND-TRUTH ACTIVITY] being performed in the image?" For object concepts: "Can you see [GROUND-TRUTH OBJECT] in the image?" Answer with only Yes or No.
S2: Visual Relationship Identification	Q1: [TRUE RELATION BETWEEN TARGET ENTITY AND GROUND-TRUTH CONCEPT]? Q2: [TRUE RELATION BETWEEN THE OTHER ENTITY AND COMPETING CONCEPT]? Answer each question with only Yes or No.
S3: Misleading Visual Cue Exclusion	Q1: [MISLEADING RELATION BETWEEN TARGET ENTITY AND COMPETING CONCEPT]? Q2: [MISLEADING RELATION BETWEEN THE OTHER ENTITY AND GROUND-TRUTH CONCEPT]? Answer each question with only Yes or No.
S4: Decision	Original multimodal input. No additional diagnostic question is provided.

Table 4: Shared prompt templates for evaluating the four stages of the diagnostic framework. S2 verifies the correct visual relationships, whereas S3 constructs misleading relationships by exchanging the concepts associated with the corresponding entities. Bracketed fields are instantiated from each word-completion item.

Here, [GROUND-TRUTH CONCEPT] denotes the visual concept corresponding to the correct completion. [TARGET ENTITY] denotes the entity queried by the original completion task, and [COMPETING CONCEPT] denotes the visually present concept that forms the competing interpretation. S2 instantiates propositions corresponding to the correct visual relationships observed in the image. S3 uses the same entities and concepts but pairs them incorrectly to form misleading visual relationships. For S1, the surface form is adapted to the semantic type of the ground-truth concept, such as an object or an activity. S4 uses the original image and incomplete caption directly, without any additional diagnostic prompt.

A.3 Cross-Modal Coverage Annotation

Figures 6–8 illustrate the three-step annotation process using the same example.

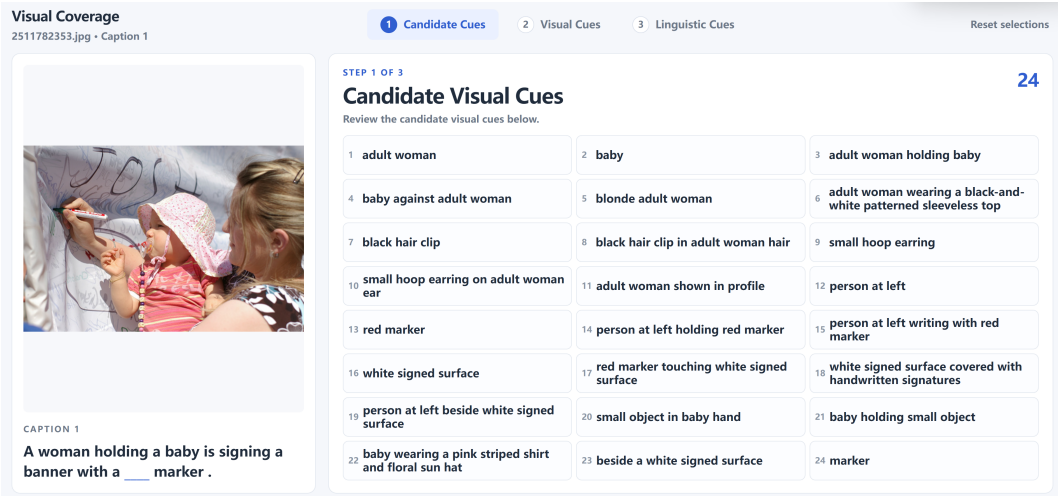


Figure 6: Construction and review of candidate visual cues. Observable entities, attributes, and relations in the example image are organized into 24 candidate visual cues.

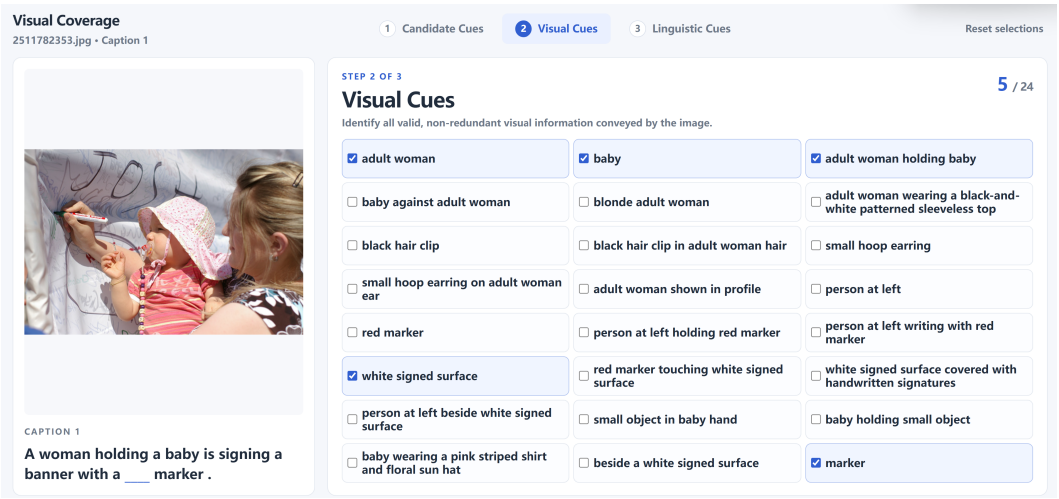


Figure 7: Selection of visual facts for the denominator. Five facts are selected from the candidate cues to form the visual-fact set $\mathcal{F}$: adult woman, baby, adult woman holding baby, white signed surface, and marker.

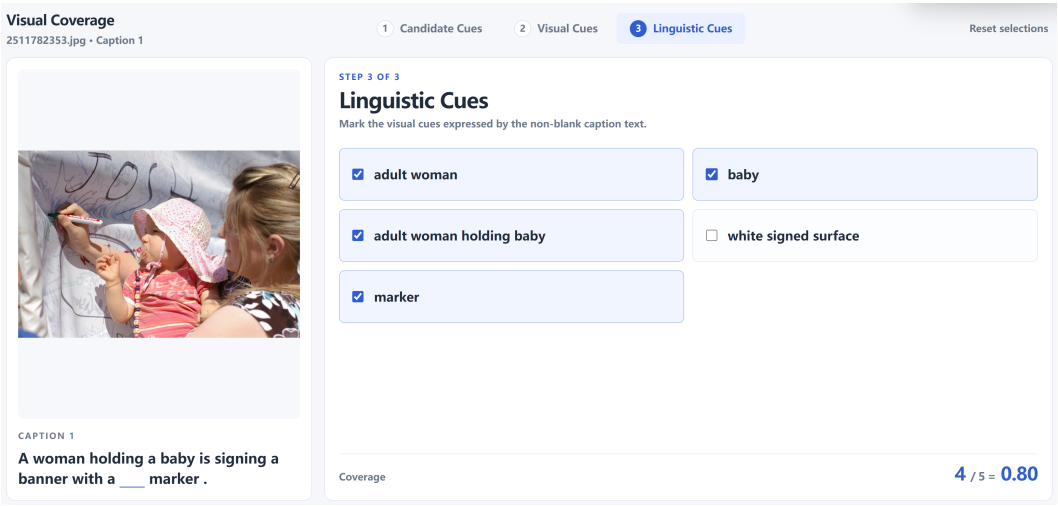


Figure 8: Identification of text-expressed facts for the numerator. Four of the five selected facts are expressed in the non-blank caption text, forming $\mathcal{K}(T)$ and yielding a cross-modal coverage of $4/5 = 0.8$.

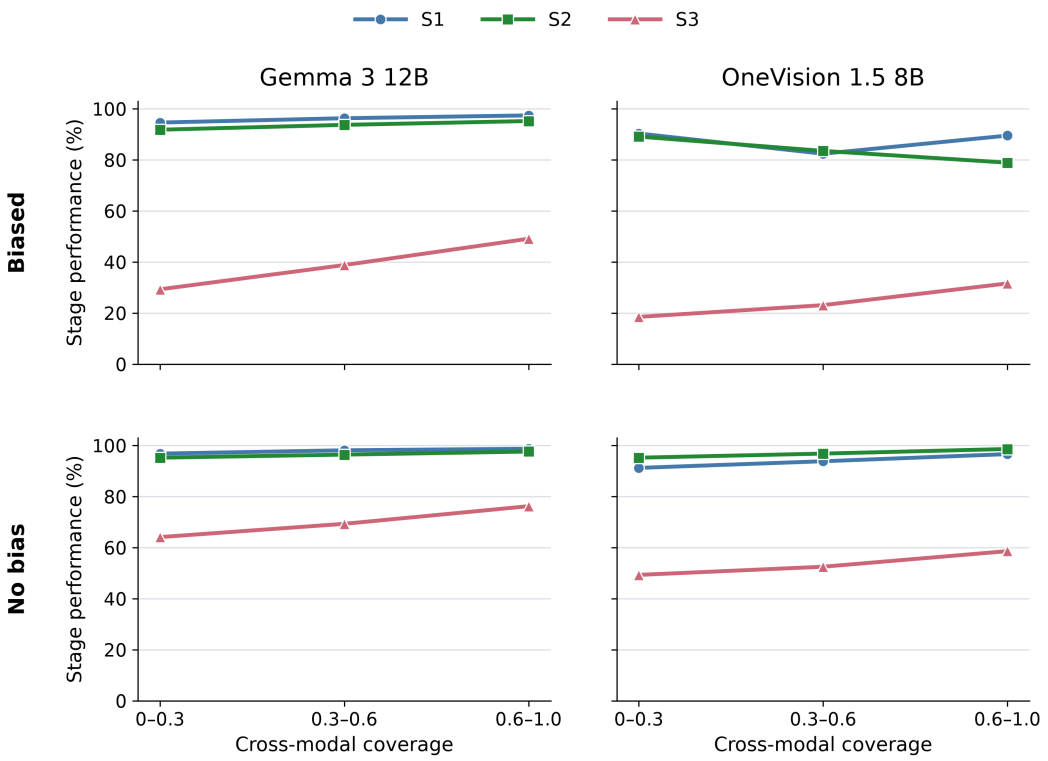


Figure 9: Additional stage-wise performance results for Gemma 3 12B and OneVision 1.5 8B across cross-modal coverage intervals. S1–S3 performance is reported separately for biased and no-bias cases. The same qualitative pattern observed in the main-text models is also present for these additional model sizes.

A.4 Additional Stage-wise Results

Figure 9 reports the same stage-wise performance analysis for Gemma 3 12B and OneVision 1.5 8B. These additional model sizes show the same qualitative pattern as the representative models reported in the main text: S1 and S2 remain relatively stable across cross-modal coverage levels, while S3 shows a clearer improvement as cross-modal coverage increases. This consistency suggests that the stage-wise observations in Figure 5 are not specific to the model sizes selected for visualization in the main text.

A.5 Prior-Counterfactual Examples

We present ten classic counterfactual cases on distinct images. For every case, the image, visual proposition, target, slot, gold answer, and coverage are fixed; only a meaning-preserving reformulation lowers the prior of the competing word. The complete five-model witness set and validation records are provided in the supplementary material.



Case 1: Qwen2.5-7B-Instruct

Original caption: A girl is smelling a mushroom that a woman is holding up to her ___.

Reformulated caption: A girl is smelling a mushroom that a woman is holding in front of her ___.

Prior of *nose*: $0.9981 \rightarrow 0.1481$

Answer: *nose* $\rightarrow$ *face*



Case 2: Qwen2.5-7B-Instruct

Original caption: A man holds up a __ while sitting in a pool of water situated on a tarp and grassy field.

Reformulated caption: While sitting in a pool of water on a tarp and grassy field, a man holds up a __.

Prior of *tube*: $0.9526 \rightarrow 0.8671$

Answer: *tube* $\rightarrow$ *child*



Case 3: Qwen2.5-7B-Instruct

Original caption: A woman in a black bikini holds a baby at the beach, while another little girl watches a __.

Reformulated caption: A little girl watches a __ at the beach, where a woman in a black bikini holds a baby.

Prior of *paddle*: $0.9399 \rightarrow 0.0067$

Answer: *paddle* $\rightarrow$ *dog*



Case 4: Qwen2.5-7B-Instruct

Original caption: A man holding a __ at the dinner table.

Reformulated caption: At the dinner table, a man is holding a __.

Prior of *corn*: $0.9963 \rightarrow 0.8846$

Answer: *corn* $\rightarrow$ *baby*



Case 5: Qwen2.5-7B-Instruct

Original caption: A man in a __ shirt and a man in an orange shirt jump in the air.

Reformulated caption: Men jump in the air, one wearing a __ shirt and the other orange.

Prior of *blue*: $> 0.9999 \rightarrow 0.8808$

Answer: *blue* $\rightarrow$ *yellow*



Case 6: Qwen2.5-7B-Instruct

Original caption: A lone red, __, and black race car is being driven by a single driver on a racetrack.

Reformulated caption: A lone race car, being driven by a single driver, is red, __, and black on a racetrack.

Prior of *Porsche*: $> 0.9999 \rightarrow 1.7 \times 10^{-7}$

Answer: *Porsche* $\rightarrow$ *white*



Case 7: Qwen2.5-7B-Instruct

Original caption: Women standing at a __ with a crowd and building in the background.

Reformulated caption: A building and a crowd appear in the background while women stand at a __.

Prior of *steps*: $0.9968 \rightarrow 0.5925$

Answer: *steps* $\rightarrow$ *podium*



Case 8: Qwen2.5-7B-Instruct

Original caption: A man playing guitar and a woman wearing a __.

Reformulated caption: A man and a woman, one playing guitar, the other wearing a __.

Prior of *red*: $> 0.9999 \rightarrow 0.7163$

Answer: *red* $\rightarrow$ *shirt*



Case 9: Qwen2.5-7B-Instruct
Original caption: A man on a boat __ looking onto the water at a boat.
Reformulated caption: A boat is near a man on a boat __, looking at the water.
Prior of *watches*: $> 0.9999 \rightarrow 7.3 \times 10^{-7}$

Answer: *watches* $\rightarrow$ *dock*



Case 10: Qwen2.5-7B-Instruct
Original caption: A man in an orange jumpsuit rests a hand on a very large reel of __ rope.
Reformulated caption: A very large reel holds __ rope, and a man in an orange jumpsuit rests a hand on it.
Prior of *nylon*: $> 0.9999 \rightarrow 6.8 \times 10^{-8}$

Answer: *nylon* $\rightarrow$ *thick*

In all ten cases, Recognition, Grounding, Relational Binding, and Exclusion remain successful before and after reformulation. The gallery is qualitative evidence for the controlled prior intervention and is not a population repair-rate estimate.