CuE: A Cue Engineering Framework for Diagnosing and Mitigating Shortcut Effects on Zero-shot MLLM Recognition

Anonymous authors
Paper under double-blind review

Abstract

Generative multimodal large language models (MLLMs) offer open-vocabulary recognition, but their decisions can depend on spurious visual or linguistic cues. We introduce CuE, a training-free framework that organizes candidate cue mechanisms, constructs target-conditioned cue hypotheses with an offline helper LLM, and performs fixed-label diagnosis before cue-matched remediation on frozen MLLMs based on a taxonomy of cue families introduced in this study. On a stratified 12-target COCO panel, we evaluate four target MLLMs from three model families. Visual diagnosis reveals recurring sensitivities: target recall is 4.7%–13.5% lower in competitor-present groups, and partial-evidence groups expose -33.0%–-20.9% recall gaps. Language-side diagnosis further shows model-dependent sensitivities: query rewrite flips 2.0%–3.6% of decisions, while language-prior probes and text interference increase miss rates by 1.3%–9.7% and 2.4%–30.1%, respectively. Cue-matched remediation improves balanced accuracy consistently across the evaluated visual settings, whereas generic debiasing baselines can overconstrain recognition and reduce recall. These findings suggest that robust zero-shot MLLM recognition requires diagnosing the relevant shortcut cue families before applying remediation, rather than relying on one-size-fits-all prompting fixes.

1 Introduction

Multimodal large language models (MLLMs) are increasingly used as open-vocabulary recognizers. Given an image and a natural-language query, they can determine the presence of a target concept without task-specific retraining. Despite their strong average recognition capabilities, achieving recognition robustness remains challenging. A model may rely on shortcut visual cues instead of target-defining features, and misleading text may override relevant visual evidence. We use *cue* to denote an observable visual or linguistic condition associated with a target. When a cue is non-causal to the target yet statistically associated with it in the data, it constitutes a *spurious correlation (SC)*. Meanwhile, once model predictions become sensitive to such a cue, we term it a *shortcut cue*. Recent studies have documented related failures, including contextual hallucination, missed detections, and text-over-vision behavior (Hosseini et al., 2026; Ye et al., 2026a; Yang et al., 2025; Cai et al., 2026). Such failures are consequential in safety-critical settings where perception errors can affect downstream planning (Xie et al., 2025; Chen et al., 2026; Hou et al., 2025).

Such robustness challenges are particularly pronounced in generative MLLMs. Supervised classifiers map images to a fixed label space, while CLIP-style vision-language models (VLMs) rank candidate class descriptions using image–text similarity scores. Generative MLLMs instead accept arbitrary natural-language queries, combine visual and textual evidence during generation, and return free-form responses. Consequently, shortcut cues can arise from multiple sources, including image content, query formulation, learned language priors, and interactions between visual and textual evidence. This broader and more heterogeneous cue space makes it substantially harder to identify cues driving a particular recognition failure and interventions appropriate for that failure.

Prior robustness studies provide important foundations but usually rely on diagnostic handles that are unavailable in zero-shot MLLM recognition. Group-aware methods exploit known SC annotations (Sagawa et al., 2020; Kirichenko et al., 2023; Idrissi et al., 2022; Ou & Zhang, 2024; Goel et al.,

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.

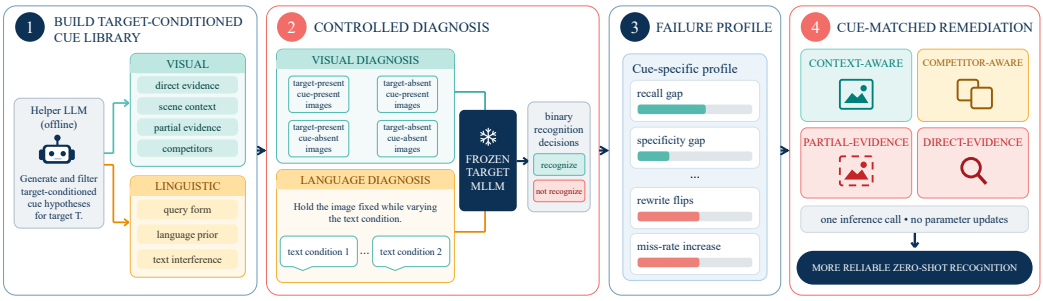


Figure 1: Overview of the proposed CuE framework.

2020), while group-free and prior-free methods infer group membership or discover hidden spurious cues from training behavior, prediction errors, captions, tags, or extracted features (Liu et al., 2021; Nam et al., 2020; Zhang et al., 2022; Kim et al., 2024; Luo et al., 2024; Yang et al., 2024; Zheng et al., 2024; Han & Zou, 2024). Related CLIP-based approaches exploit prompt selection, embedding intervention, or text-guided bias editing (Varma et al., 2024; Ye et al., 2026b; Molahasani et al., 2025; Lu et al., 2025; Yalavarthi et al., 2026). Recent MLLM studies address related failures through contrastive decoding, counterfactual debiasing, neuron-level intervention, routing, or perturbation-based fine-tuning (Cai et al., 2026; Leng et al., 2023; Wu et al., 2025; An et al., 2026). Despite these advances, existing approaches generally either require training data or internal model access, assume predefined bias descriptions or groups, or target a particular failure mechanism. They therefore provide limited support for systematically discovering spurious cues that drive recognition failures.

This paper argues that **the broader and more heterogeneous shortcut space of generative MLLMs calls for a taxonomy-guided diagnosis-before-remediation procedure**. Instead of applying a generic debiasing intervention, we first identify cue families that affect recognition for a given target and then select a remediation strategy accordingly. This leads to three key questions: (1) which cue families affect recognition performance for a given target? (2) what kind of failure patterns does this sensitivity reveal? and (3) can remediation tailored to the diagnosed cue produce more reliable improvements than a generic intervention?

We attempt to address these questions with **CuE**, a training-free framework that organizes candidate cues into an extensible taxonomy and uses an offline helper LLM to construct target-conditioned cue hypotheses. An overview of the framework is shown in Figure 1, where these hypotheses are evaluated through controlled diagnostic evaluations on frozen MLLMs. Visual diagnostics compare cue-present and cue-absent image groups for each target, while linguistic diagnostics vary the textual condition while holding the image fixed. The resulting diagnosis then guides cue-matched remediation, allowing the intervention to target the identified failures rather than imposing a one-size-fits-all intervention.

We evaluate CuE on a stratified 12-target COCO panel with four MLLMs. The diagnosis reveals recurring visual sensitivities to competitor, context, and partial-evidence cues, with target recall gaps of 4.7%–13.5% for competitor groups and -33.0%–-20.9% for partial-evidence groups. Linguistic diagnostics reveal model-dependent sensitivities: query rewrites change 2.0%–3.6% of decisions, while language-prior probes and text interference increase miss rates by 1.3%–9.7% and 2.4%–30.1%, respectively. Cue-matched remediation improves balanced accuracy across the evaluated visual model–cue settings, whereas generic debiasing baselines can over-constrain recognition and reduce recall. Together, these results support diagnosis as an important step in selecting remediation strategies for zero-shot generative MLLM recognition.

Contributions.

- We formalize cue-driven shortcut sensitivity for zero-shot MLLM recognition by distinguishing data-level cue associations from model sensitivity under a fixed ground-truth target.
- We introduce an extensible taxonomy of spurious cues as well as an offline helper-LLM procedure for constructing target-conditioned cue libraries, enabling systematic shortcut cue diagnosis.
- We develop controlled visual and linguistic diagnostics that isolate cue conditions while keeping the ground-truth target fixed, as well as use them to characterize recurring and model-dependent shortcut sensitivities across multiple MLLM families.

- We introduce cue-matched remediation based on the diagnosed failure patterns while showing that it provides more consistent improvements across the evaluated settings than a taxonomy-agnostic prompting strategy without requiring target-dataset fine-tuning or auxiliary modules.

2 Related Work

Robustness against spurious correlations via groups and inferred biases. Robustness against SCs is commonly studied by partitioning data into groups defined by target labels and spurious attributes. Group-aware methods assume access to such group annotations and improve worst-group performance through robust objectives, reweighting, data augmentation, or retraining (Sagawa et al., 2020; Kirichenko et al., 2023; Idrissi et al., 2022; Ou & Zhang, 2024; Goel et al., 2020; Qiu et al., 2023). Group-free methods reduce the need for explicit group annotations by exploiting prediction errors, bias-learning dynamics, or inferred group structure (Liu et al., 2021; Nam et al., 2020; Zhang et al., 2022; Yang et al., 2024; Zheng et al., 2024; Han & Zou, 2024). Language-assisted discovery methods further use semantic descriptions to identify candidate biases or subpopulation structure (Kim et al., 2024; Luo et al., 2024); other approaches discover biases through unsupervised analysis (Guimard et al., 2025). Collectively, these studies establish the importance of identifying structured subpopulations in which model performance differs, but they primarily consider supervised classifiers with a fixed label space and access to training data, annotations, or learned representations.

Language-guided bias discovery and intervention in VLMs. For CLIP-style VLMs, robustness methods can operate on candidate prompts or the shared image–text embedding space. SAGE selects prompts based on class separation without model updates or external annotations (Ye et al., 2026b). Text-guided methods use descriptions of spurious attributes to project or translate embeddings and reduce unwanted associations (Molahasani et al., 2025; Lu et al., 2025). RAVL discovers spurious local image features through region-level clustering and mitigates their effects with a region-aware fine-tuning loss (Varma et al., 2024). DIAL uses sparse autoencoders to identify spurious feature directions and remove their contribution from affected representations (Yalavarthi et al., 2026). However, these methods rely on fixed-label or embedding-based intervention handles, whereas CuE diagnoses visual and linguistic cues through controlled input changes without accessing or modifying MLLM representations.

Shortcut diagnosis and mitigation in generative MLLMs. Recent work has begun to investigate spurious cues and shortcut behavior directly in generative MLLMs, but existing studies address different components or manifestations of the problem. MM-SpuBench introduces a human-verified benchmark covering nine categories of multimodal spurious bias and evaluates models primarily through average VQA accuracy (Ye et al., 2026a). SpurLens uses GPT-4 and open-set object detectors to discover visual cues and measures their effects on object recognition and hallucination (Hosseini et al., 2026). SpuriVerse derives correlations from real-world VQA errors and synthetic counterfactual examples, and then studies whether fine-tuning improves generalization to unseen correlations (Yang et al., 2025). Modality Interference diagnoses model susceptibility to cross-modal interference and mitigates its effects through perturbation-based augmentation and consistency-regularized fine-tuning (Cai et al., 2026). Other methods target particular manifestations, such as object hallucination or spurious visual reliance, through visual contrastive decoding or counterfactual inference (Leng et al., 2023; Wu et al., 2025).

Together, these studies establish that generative MLLMs can exhibit sensitivity to both visual and linguistic factors. However, they leave a methodological gap in systematic cue diagnosis and remediation for frozen MLLM recognition. Existing evaluations typically report aggregate accuracy, likelihood, or overall effects under a particular perturbation, while corresponding interventions are often generic, modality-specific, or training-based. More fundamentally, existing approaches do not systematically connect a candidate cue to its effect on recognition, the resulting failure patterns, and the corresponding remediation. CuE addresses this gap by systematically evaluating how each candidate cue affects recognition, identifying the resulting cue-specific failure pattern, and selecting a matched remediation strategy within a unified framework.

3 CUE: A CUE ENGINEERING FRAMEWORK

CuE is a training-free framework for diagnosing and remediating recognition failures induced by shortcut cues in zero-shot MLLMs. An overview of the framework is presented in Figure 1. The framework follows a diagnosis-before-remediation procedure. First, it organizes candidate visual and linguistic cues into an extensible taxonomy, constructs a target-conditioned cue-hypothesis library, and evaluates whether a frozen target MLLM is sensitive to these cues under controlled conditions. The resulting failure profile then guides the construction of cue-conditioned prompts for remediation. Each cue family is evaluated separately to facilitate understanding of the way that the hypothesized cue relates to the observed failure pattern and the corresponding intervention.

3.1 PROBLEM SETUP: CONTROLLED CUE VARIATION

Let X denote an image, T a queried target concept, and $Y_T \in \{0, 1\}$ its ground-truth presence label. Given an image x and query $q(T)$, a frozen MLLM produces a response that is parsed into a binary decision $\hat{Y}_T$. Let $C \in \{0, 1\}$ indicate whether a specified visual cue is present. Data-level cue association can be expressed as:

$$P(C = 1 \mid Y_T = 1) \gg_{\rho,\delta} P(C = 1 \mid Y_T = 0),$$

where $a \gg_{\rho,\delta} b$ denotes $(a + \epsilon)/(b + \epsilon) \geq \rho$ and $a - b \geq \delta$, with $\rho > 1$, $\delta > 0$, and $\epsilon > 0$. This relation requires both relative and absolute differences to be non-trivial. Here, P refers to the underlying data distribution; the deliberately balanced diagnostic groups are used to measure model sensitivity, not to estimate natural cue–target co-occurrence.

Association alone does not establish model reliance on a cue. CuE therefore compares predictions across cue conditions while conditioning on the same target label. For visual diagnosis, we define:

$$\mathrm{Recall}_{gap} = P(\hat{Y}_T = 1 \mid Y_T = 1, C = 0) - P(\hat{Y}_T = 1 \mid Y_T = 1, C = 1), \tag{1}$$

$$\mathrm{Spec}_{gap} = P(\hat{Y}_T = 0 \mid Y_T = 0, C = 0) - P(\hat{Y}_T = 0 \mid Y_T = 0, C = 1). \tag{2}$$

A positive Recall_{gap} indicates lower recall when the cue is present, whereas a positive Spec_{gap} indicates more false positives when the cue is present. Together, these quantities characterize model sensitivity to the cue under the controlled evaluation conditions. For linguistic diagnosis, the image is instead held fixed while the relevant textual condition is varied.

3.2 CUE-FAMILY TAXONOMY

To systematically search for potential shortcut cues, CuE organizes candidate visual and linguistic conditions into seven cue families. The visual families are *direct evidence*, *scene context*, *partial evidence*, and *competitors*, while the linguistic families are *query form*, *language prior*, and *text interference*. The taxonomy is designed to cover distinct ways in which visual or linguistic conditions can affect open-vocabulary recognition while remaining extensible to additional cue families.

Importantly, membership in a cue family does not by itself imply that a cue is spurious or that the model relies on it. The taxonomy specifies candidate conditions to be tested. Shortcut sensitivity is established only through the controlled diagnosis described in Section 3.4.

3.2.1 VISUAL CUE FAMILIES

Direct evidence refers to intrinsic visual features that directly support identification of the target concept. For example, the characteristic shape and structural configuration of an object can provide target-defining evidence for its recognition. Direct evidence is not treated as a nuisance cue. Instead, it serves as a reference condition for examining whether explicitly directing the model toward target-defining evidence improves recognition or imposes unnecessarily restrictive evidence requirements.

Scene context refers to environmental, spatial, or situational information that commonly accompanies a target but is not part of the target itself. Examples include roads surrounding cars or branches surrounding birds. A scene-context cue is target-typical and target-predictive: $P(C = 1 \mid Y_T = 1) \gg_{\rho,\delta} P(C = 1 \mid Y_T = 0)$, $P(Y_T = 1 \mid C = 1) \gg_{\rho,\delta} P(Y_T = 1 \mid C = 0)$. Yet the cue is not target-defining. The context can occur without the target, and the target can appear outside its usual

context. Diagnosis therefore examines misses on $(Y_T = 1, C = 0)$ images and false positives on $(Y_T = 0, C = 1)$ images.

Partial evidence refers to local visual parts, attributes, or closely associated objects that provide suggestive evidence for a target but are insufficient to establish its presence. Examples include wheels as partial evidence for a bicycle or a cabin for a train. Such cues may be predictive, with $P_{\text{data}}(Y_T = 1 \mid C = 1) \gg_{\rho,\delta} P_{\text{data}}(Y_T = 0 \mid C = 1) > 0$, yet they can also occur without the target. This family is distinguished from scene context by its more direct local or structural relationship to the target, and from direct evidence by its insufficiency for target identification. Diagnosis therefore considers both false recognition induced by partial cues and missed recognition when such familiar cues are absent.

Competitors refer to distinct objects that share visual, functional, or contextual characteristics with the queried target and may therefore compete for recognition. Examples include a snowboard for skis or a truck for a bus. Unlike scene context, a competitor can support an alternative object interpretation. Competitor cues can produce competitive suppression when the presence of a competing object reduces recognition of a target that is actually present. We therefore examine both target-positive examples with and without competitors and target-negative examples containing competitors to distinguish suppression from target–competitor confusion.

3.2.2 Linguistic Cue Families

Query form captures variations in the linguistic realization of the same target concept, including synonyms, paraphrases, and answer-format variants. These variations preserve the intended recognition target while changing how the query is expressed. Query-form sensitivity is measured by decision variation across equivalent queries. Sensitivity to query form indicates that recognition decisions may depend on the formulation of the query rather than solely on the visual evidence.

Language prior refers to textual conditions that alter the plausibility of the target according to learned world knowledge, common co-occurrence patterns, stereotyped associations, or other language-derived expectations. By varying such conditions given the same image, we can test whether linguistic plausibility changes the recognition decision even when the underlying visual evidence is unchanged.

Text interference refers to additional textual content that is irrelevant to, contradictory to, or suggestive about the queried target. Examples include unrelated statements, contradictory facts, user judgments, competitor suggestions, and biased instructions. Unlike query-form variation, which preserves the semantic content of the query, text interference introduces additional information that may compete with or override visual evidence. It therefore provides a direct test of whether non-essential textual information alters recognition decisions.

3.3 Offline Helper-LLM Cue Library Construction

The taxonomy defines the candidate cue space but does not specify actual cues that are relevant to a particular target. CuE therefore uses an offline helper LLM to construct a target-conditioned cue library. For each target T, the helper generates candidate hypotheses for the seven cue families:

$$\mathcal{H}(T) = \big(\mathcal{H}_{dir}(T), \mathcal{H}_{ctx}(T), \mathcal{H}_{part}(T), \mathcal{H}_{comp}(T), \mathcal{H}_{query}(T), \mathcal{H}_{prior}(T), \mathcal{H}_{text}(T)\big).$$

Each hypothesis specifies a concrete visual or linguistic condition that can be instantiated in the subsequent diagnostic evaluation. For example, for the target *bus*, the library may contain roads or bus stops as scene-context hypotheses and trucks as competitor hypotheses.

Generated candidates are subsequently filtered using schema checks, deduplication, and family-specific semantic criteria to ensure the validity of the candidate library. For example, context cues must be distinguishable from the target and able to occur without it; partial-evidence cues must be suggestive but insufficient; competitors must represent distinct alternative objects; and query rewrites must preserve the target semantics. Candidates that fail a required criterion are rejected. If no candidate survives, a bounded recovery step regenerates candidates and applies the same checks. These procedures reduce the number of invalid hypotheses.

The resulting library is constructed once per target and reused across target MLLMs. The helper LLM only proposes and filters hypotheses, while cue sensitivity is determined exclusively from the

cue-controlled diagnosis of the frozen target MLLM. Appendix B provides the generation, filtering, and remediation templates.

3.4 CUE-CONTROLLED DIAGNOSIS

For each target T and cue family f, CuE instantiates diagnostic comparisons from the retained hypotheses $\mathcal{H}_f(T)$. For visual cue families, CuE constructs grouped evaluations over cue-present and cue-absent images. We measure cue-dependent recall gaps by comparing positive groups $(Y_T = 1, C = 1)$ and $(Y_T = 1, C = 0)$, while measuring cue-induced specificity gaps by comparing negative groups $(Y_T = 0, C = 1)$ and $(Y_T = 0, C = 0)$. The interpretation depends on the cue family. For linguistic cue families, CuE holds the image fixed while varying the textual condition, allowing us to assess whether recognition decisions change with query wording or language-side information despite unchanged visual evidence.

The diagnosis serves three purposes. First, it tests whether cue hypotheses expose meaningful model brittleness. Second, it tests whether the target MLLM suffers from shortcut-induced failures under controlled cue variation. Finally, it diagnoses failure modes of each cue family.

3.5 CUE-CONDITIONED PROMPT REMEDIATION

The diagnosed failure profile guides the construction of cue-conditioned prompts for remediation. For each diagnosed visual cue family, CuE augments the base recognition query with a family-specific instruction that targets the identified failure pattern.

Context-aware prompts clarify that surrounding context may support recognition but does not establish target presence; partial-evidence prompts distinguish suggestive parts or associated objects from sufficient target evidence; competitor-aware prompts instruct the model to distinguish the queried target from visually similar alternatives; and direct-evidence prompts encourage verification of target-defining visual features.

Each cue-conditioned prompt is evaluated against the base query on the same images using balanced accuracy and group-specific recall and specificity. Improvements are interpreted together with the affected groups to determine whether an intervention alleviates the diagnosed failure pattern rather than simply shifting the model toward positive or negative predictions. Generic direct-evidence prompting is also evaluated as a taxonomy-agnostic baseline to examine whether diagnosis-guided interventions provide more targeted improvements.

The remediation operates independently for each cue family and does not require parameter updates or auxiliary modules. Cue generation and diagnosis are performed offline, while remediation consists of replacing the base recognition query with the corresponding cue-conditioned prompt in a single inference call to the same frozen MLLM.

4 EXPERIMENTS

This section evaluates CuE in two stages: cue-controlled diagnosis and cue-conditioned remediation. Concretely, we aim to answer: (1) Can target-conditioned cue hypotheses reveal failures that are hidden by aggregate accuracy? (2) Which cue families induce the largest recall or specificity gaps? (3) Do cue-matched prompts improve robustness for the corresponding cue-controlled evaluations? (4) Can generic debiasing promptings provide reliable remediation across different cue families?

4.1 EXPERIMENTAL DETAILS

4.1.1 EVALUATION DATASET CONSTRUCTION

We evaluate on the COCO train split (Lin et al., 2014) using a deliberately stratified panel of 12 targets: animals (*bird*, *dog*, *giraffe*), vehicles (*car*, *bus*, *train*), indoor objects (*couch*, *chair*, *vase*), and wearable or sports concepts (*tie*, *skis*, *tennis racket*). For each target, we construct four evaluation sets. The *general evaluation set*, used for base recognition and language-side evaluations, contains 1000 positive images, 500 hard negatives, and 500 generic negatives. Hard negatives are target-absent images containing resolved competitors, scene-context cues, or partial-evidence cues matched

to COCO categories, while generic negatives are target-absent images sampled without requiring helper-generated cue presence. The *context evaluation set* contains four groups, $(Y_T = 1, C = 0)$, $(Y_T = 0, C = 0)$, $(Y_T = 1, C = 1)$, and $(Y_T = 0, C = 1)$, where C denotes resolved scene-context cue presence; each group contains 500 samples. The *competitor evaluation set* and *partial-evidence evaluation set* use the same four-group construction, with C denoting competitor presence and partial-evidence presence, respectively.

4.1.2 Experimental Details

We use Qwen3.5-9B (Team, 2026) as the offline helper LLM. It constructs and filters one library per target, which is then reused across target MLLMs. We evaluate Qwen2.5-VL-7B-Instruct (Qwen2.5) (Team, 2025a), Qwen3-VL-8B-Instruct (Qwen3) (Team, 2025b), Gemma-4-12B (Gemma-4) (Team et al., 2026), and GLM-4.6V-Flash (GLM-4.6V) (Team et al., 2025). For visual remediation, each cue family is evaluated with cue-conditioned prompts to examine the trade-off between reducing cue-induced false positives and preserving true-positive recognition.

4.1.3 Metrics

We report recall, $\mathrm{Recall} = P(\hat{Y}_T = 1 \mid Y_T = 1)$, specificity, $\mathrm{Specificity} = P(\hat{Y}_T = 0 \mid Y_T = 0)$, and balanced accuracy, $\mathrm{BAcc} = (\mathrm{Recall} + \mathrm{Specificity})/2$. For query-form evaluation, we report the rewrite flip rate, defined as the fraction of samples for which target-preserving query rewrites produce different binary decisions. For language-prior and text-interference evaluations, we report the miss-rate increase relative to the base query $\mathrm{Miss}_{gap} = \mathrm{Miss}_{q_{cue}} - \mathrm{Miss}_{q_{base}}$, where $\mathrm{Miss}_q = P(\hat{Y}_T = 0 \mid Y_T = 1, q)$. For grouped visual evaluations, we report $\mathrm{Recall}_{C=0} = P(\hat{Y}_T = 1 \mid Y_T = 1, C = 0)$, $\mathrm{Recall}_{C=1} = P(\hat{Y}_T = 1 \mid Y_T = 1, C = 1)$, $\mathrm{Spec}_{C=0} = P(\hat{Y}_T = 0 \mid Y_T = 0, C = 0)$, and $\mathrm{Spec}_{C=1} = P(\hat{Y}_T = 0 \mid Y_T = 0, C = 1)$, as well as Recall_{gap} and Spec_{gap} defined in Section 3.1. Unless stated otherwise, results are averaged over targets.

4.2 Diagnostic Evaluation

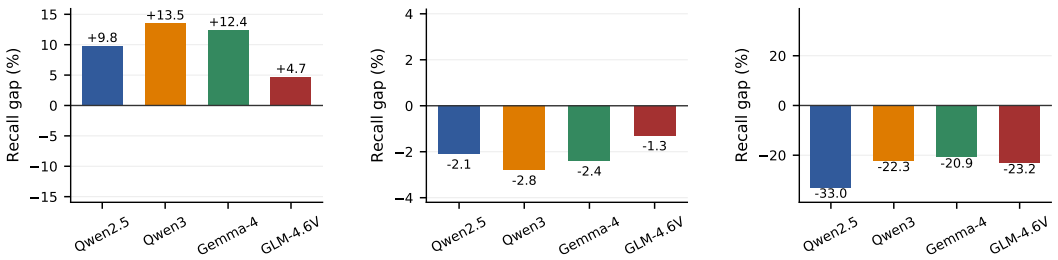
Aggregate accuracy hides cue-specific failures. The base-query balanced accuracy on the general evaluation set is 92.9%, 95.1%, 93.8%, and 95.5% for Qwen2.5, Qwen3, Gemma-4, and GLM-4.6V, respectively. These strong aggregate scores indicate capable zero-shot recognition on average, making the cue-specific gaps more informative than uniformly poor recognition results.

Visual cue families reveal distinct failure patterns. Figure 2 summarizes grouped visual evaluation, with full results in Table 3. Competitor co-presence consistently suppresses target recognition, with recall gaps of 4.7%–13.5%, while specificity gaps of 5.0%–14.3% indicate target–competitor confusion on competitor-present negatives. Context groups produce smaller but consistent negative recall gaps, indicating degraded recognition when the usual scene context is absent. Partial-evidence groups show the strongest recall asymmetry (-33.0% to -20.9%) and also reduce specificity by 10.3%–17.7%, indicating false recognition from insufficient visual evidence.

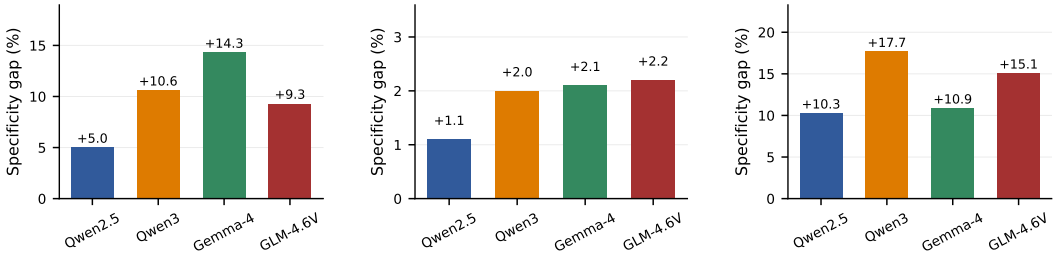
A model can therefore achieve high overall accuracy while remaining systematically sensitive to specific cues that occur unevenly across examples. This motivates evaluating recognition under controlled cue conditions, where the relevant cue is explicitly varied. Such cue-controlled evaluation can expose failure patterns that are diluted in aggregate metrics and provide a more informative basis for diagnosing and subsequently remediating shortcut reliance.

Language-side cue families reveal brittleness even when the image is fixed. Figure 3 shows model-dependent sensitivity to linguistic cues. Target-preserving rewrites change 2.0%–3.6% of decisions. Language-prior probes increase missed detections by 1.3%–9.7%, while text interference produces a larger and more variable increase of 2.4%–30.1%. These effects occur with the image fixed, indicating that changes in the textual condition can substantially alter recognition, particularly through increased missed detections.

Text-interference subtypes differ substantially. The subtype heatmap in Figure 3d further reveals substantial model- and subtype-dependent sensitivity. Contradictory and irrelevant facts produce large increases in missed detections for Qwen2.5 (41.4% and 53.3%) and Qwen3 (42.4% and 22.0%), respectively. User judgments also have pronounced effects on Qwen2.5 and GLM-4.6V, increasing

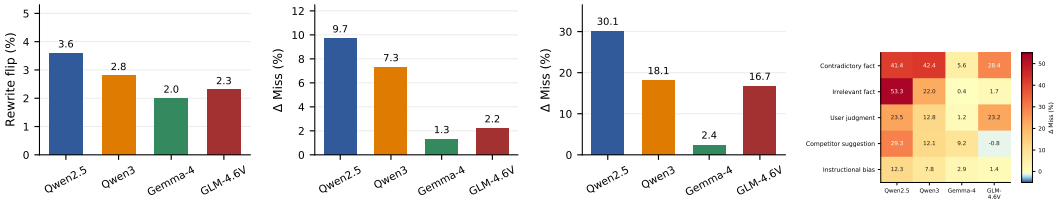


(a) Competitor recall. (b) Context recall. (c) Partial-evidence recall.



(d) Competitor specificity. (e) Context specificity. (f) Partial-evidence specificity.

Figure 2: Grouped visual evaluation results.



(a) Query rewrites. (b) Language prior. (c) Text interference. (d) Interference subtypes.

Figure 3: Language-side diagnosis with the image held fixed. The effects are model-dependent and are dominated by missed detections.

missed detections by 23.5% and 23.2%, respectively. In contrast, Gemma-4 is relatively less affected by these subtypes, although competitor suggestions still increase missed detections by 9.2%. These differences suggest that language-side sensitivity is not uniform across models or interference types, motivating more fine-grained cue diagnosis beyond a single aggregate measure of text sensitivity.

4.3 Remediation Evaluation

Cue-conditioned prompt remediation benefits are cue-specific. Table 1 reports the performance improvements of CuE over the base query. CuE improves balanced accuracy across all evaluated model–cue settings, with gains of 0.73%–2.65%. The improvements also align with the diagnosed failure patterns: competitor remediation improves performance on competitor-present groups, while context and partial-evidence remediation primarily recovers recognition on cue-absent target-positive groups and improves specificity on cue-present negatives. These results show that diagnosing the affected cue family can guide targeted remediation without model retraining.

Cue matching is more consistent than taxonomy-agnostic prompting. Table 2 reports ΔBAcc relative to the base query across 12 model–cue settings, comparing cue-conditioned prompt remediation with three generic debiasing baselines adapted from MM-SpuBench (Ye et al., 2026a), including describe-then-decide, explain-then-decide, and generic-debiasing prompts. Full prompts are provided in Appendix A.3. CuE improves balanced accuracy in all 12 settings, compared with 2/12 for generic debiasing, 5/12 for explain-then-decide, and 2/12 for describe-then-decide.

Although the baselines occasionally achieve larger gains in individual settings, these improvements can reflect different recall–specificity trade-offs rather than consistently addressing the diagnosed cue (see full results in Table 8). For example, on the Qwen2.5 context evaluation, describe-then-decide improves BAcc by 1.50%, larger than 0.73% for CuE. However, this gain comes with recall increases

Table 1: Changes in metrics (%) after cue-conditioned remediation relative to the base query.

Evaluation	Model	ΔBAcc	$\Delta\text{Recall}_{C=0}$	$\Delta\text{Recall}_{C=1}$	$\Delta\text{Spec}_{C=0}$	$\Delta\text{Spec}_{C=1}$
Competitor	Qwen2.5	+0.92	-0.1	+2.3	-0.1	+1.6
Competitor	Qwen3	+1.75	+0.1	+2.7	-0.2	+4.4
Competitor	Gemma-4	+2.65	-0.8	+3.4	+0.5	+7.5
Competitor	GLM-4.6V	+1.35	+0.8	+0.3	+0.4	+3.9
Context	Qwen2.5	+0.73	+2.2	+1.2	-0.2	-0.3
Context	Qwen3	+1.40	+3.2	+2.8	-0.1	-0.3
Context	Gemma-4	+1.70	+3.3	+2.1	+0.3	+1.1
Context	GLM-4.6V	+1.05	+2.5	+1.9	0.0	-0.2
Partial evidence	Qwen2.5	+0.73	-0.6	0.0	+0.3	+3.2
Partial evidence	Qwen3	+1.88	+2.0	+0.3	-0.3	+5.5
Partial evidence	Gemma-4	+2.25	+3.5	0.0	+1.1	+4.4
Partial evidence	GLM-4.6V	+2.25	+5.7	-0.9	+1.3	+2.9

Table 2: $\Delta\text{BAcc}(\%)$ relative to the base query.

Cue family	Model	CuE	Describe	Explain	Generic debiasing
Competitor	Qwen2.5	+0.92	+0.40	+0.05	-0.28
Competitor	Qwen3	+1.75	-0.13	0.00	-0.15
Competitor	Gemma-4	+2.65	-0.93	-2.12	-0.28
Competitor	GLM-4.6V	+1.35	-0.35	+2.20	+0.25
Context	Qwen2.5	+0.73	+1.50	+0.55	-0.07
Context	Qwen3	+1.40	-0.25	+0.10	0.00
Context	Gemma-4	+1.70	-1.00	-0.27	-0.22
Context	GLM-4.6V	+1.05	-1.10	-0.78	+0.20
Partial evidence	Qwen2.5	+0.73	-2.00	-1.23	-0.15
Partial evidence	Qwen3	+1.88	-1.08	-0.68	-0.17
Partial evidence	Gemma-4	+2.25	-1.05	-0.55	-0.62
Partial evidence	GLM-4.6V	+2.25	-0.27	+0.50	-0.22

of 5.6% and 4.4% for the cue-absent and cue-present groups, respectively, accompanied by specificity decreases of 1.5% and 2.5%. This pattern indicates a general shift toward predicting "Yes", rather than targeted correction of the diagnosed cue-related errors. In contrast, CuE increases recall by 2.2% and 1.2% while inducing only negligible changes in specificity. Overall, generic prompts can sometimes improve BAcc by broadly changing the model's prediction behavior, but their effects vary across settings. CuE uses the diagnosed cue to guide remediation, leading to more consistent improvements with smaller changes in the recall–specificity balance.

5 Conclusion

We introduced CuE, a training-free cue engineering framework for diagnosing and remediating shortcut reliance in zero-shot MLLM recognition. CuE organizes visual and linguistic shortcut mechanisms into an extensible cue-family taxonomy, uses an offline helper LLM to construct target-conditioned hypotheses, and evaluates frozen target MLLMs under fixed ground-truth target labels before selecting a family-specific intervention.

Across four MLLMs on a stratified 12-target COCO panel, visual diagnosis reveals recurring yet distinct failure profiles associated with competitor, context, and partial-evidence cues. Language-side sensitivity is model-dependent: target-preserving rewrites, language priors, and text interference can alter recognition even when the image remains fixed. Cue-matched remediation improves balanced accuracy across all evaluated visual model–cue settings and yields more consistent gains than taxonomy-agnostic prompting baselines, while generic debiasing prompts can over-constrain recognition and reduce recall. These results support a diagnosis-before-remediation approach.

AI USE STATEMENT

In this work, we used generative AI tools to clean and reformat the dataset, assist in designing experiments, provide feedback on research methodology, assist in implementing methods, and help interpret results.

Additionally, we used generative AI tools to summarize and analyze existing literature, polish the research paper to improve readability.

As part of the methodological design, generative AI tools were used in this work as an offline helper LLM to propose and filter target-conditioned cue hypotheses and to assist with prompt constructions described in Sections 3.3 and 3.5. The generated cue hypotheses were validated by human annotators (see Appendix A.2 for details).

We have not used generative AI tools to generate synthetic data sets, assist with translation, or design research methodology.

The following uses of generative AI were not applicable to this work: developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting with proofs, or proposing or refining hypotheses for theories.

The authors reviewed and validated all AI-assisted outputs, experimental procedures, analyses, and reported results. Generative AI tools did not replace author judgment in selecting the methodology, interpreting the findings, or making final claims. The authors take responsibility for the final content of this paper, including any text, claims, code, or artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

Methodological details, including the full prompt templates, are provided in Appendix B. Experimental details are provided in Section 4.1 and Appendix A. The computational environment is illustrated in Appendix D.

REFERENCES

Na Min An, Yoonna Jang, Yusuke Hirota, Ryo Hachiuma, Isabelle Augenstein, and Hyunjung Shim. Interpretable debiasing of vision-language models for social fairness, 2026. URL https://arxiv.org/abs/2602.24014.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models, 2026. URL https://arxiv.org/abs/2505.19616.

Yifei Chen, Chenqian Le, Jiayi Cheng, and Xupeng Chen. Vehanchor: Metadata-free metric scale recovery from vehicle cues in aerial imagery, 2026. URL https://arxiv.org/abs/2603.04277.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

Quentin Guimard, Moreno D'Incà, Massimiliano Mancini, and Elisa Ricci. Classifier-to-bias: Toward unsupervised automatic bias detection for visual classifiers, 2025. URL https://arxiv.org/abs/2504.20902.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference, 2024. URL https://arxiv.org/abs/2404.13815.

Parsa Hosseini, Sumit Nawathe, Mazda Moayeri, Sriram Balasubramanian, and Soheil Feizi. Spurlens: Automatic detection of spurious cues in multimodal llms, 2026. URL https://arxiv.org/abs/2503.08884.

Kaiyuan Hou, Minghui Zhao, Lilin Xu, Yuang Fan, and Xiaofan Jiang. Tdbench: A benchmark for top-down image understanding with reliability analysis of vision-language models, 2025. URL https://arxiv.org/abs/2504.03748.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11082–11092. IEEE, 2024.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2023. URL https://arxiv.org/abs/2204.02937.

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, 2023. URL https://arxiv.org/abs/2311.16922.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International conference on machine learning*, pp. 6781–6792. PMLR, 2021.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Mitigating spurious correlations in zero-shot multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UsRKFYR4lM.

Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pp. 235–252. Springer, 2024.

Mahdiyar Molahasani, Azadeh Motamedi, Michael Greenspan, Il-Min Kim, and Ali Etemad. Prism: Reducing spurious implicit biases in vision-language models with llm-guided embedding projection, 2025. URL https://arxiv.org/abs/2507.08979.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier, 2020. URL https://arxiv.org/abs/2007.02561.

Ziquan Ou and Zijun Zhang. Coda: A correlation-oriented disentanglement and augmentation modeling scheme for better resisting subpopulation shifts. *Advances in Neural Information Processing Systems*, 37:102587–102610, 2024.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pp. 28448–28467. PMLR, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL https://arxiv.org/abs/1911.08731.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Qwen2.5-VL Team. Qwen2.5-vl technical report, 2025a. URL https://arxiv.org/abs/2502.13923.

Qwen3-VL Team. Qwen3-vl technical report, 2025b. URL https://arxiv.org/abs/2511.21631.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models, 2024. URL https://arxiv.org/abs/2411.04097.

Zichen Wu, Hsiu-Yuan Huang, and Yunfang Wu. Beyond spurious signals: Debiasing multimodal large language models via counterfactual inference and adaptive expert routing, 2025. URL https://arxiv.org/abs/2509.15361.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives, 2025. URL https://arxiv.org/abs/2501.04003.

Bharat Chandra Yalavarthi, Nalini K. Ratha, and Venu Govindaraju. Label-free mitigation of spurious correlations in VLMs using sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NHOLsaHuFv.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Z. Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations?, 2025. URL https://arxiv.org/abs/2506.18322.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias, 2024. URL https://arxiv.org/abs/2305.18761.

Wenqian Ye, Bohan Liu, Guangtao Zheng, Di Wang, Xu Cao, Yunsheng Ma, Bolin Lai, James M. Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 2854–2865. ACM, April 2026a. doi: 10.1145/3770854.3785678. URL http://dx.doi.org/10.1145/3770854.3785678.

Wenqian Ye, Di Wang, Guangtao Zheng, Bohan Liu, and Aidong Zhang. Sage: Spuriousness-aware guided prompt exploration for mitigating multimodal bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27809–27817, 2026b.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Spuriousness-aware meta-learning for learning robust classifiers. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4524–4535, 2024.

A Additional Evaluation Details

A.1 Full Visual Diagnostic Results

Table 3 reports the full results of the grouped visual diagnosis. Values are macro-averaged across the 12 evaluation targets. For each cue family, we compare performance between cue-absent ($C = 0$) and cue-present ($C = 1$) groups separately for target-positive and target-negative examples. The corresponding recall and specificity gaps therefore indicate whether a cue is associated with missed target detections, false positives, or both.

The three cue families exhibit distinct failure patterns across models. For competitor cues, all four models show lower recall when the competitor is present, with recall gaps of 4.7%–13.5%. Specificity also decreases for cue-present negatives, with gaps of 5.0%–14.3%, indicating that competitor presence affects both target detection and rejection of negative examples. Context cues produce substantially smaller recall gaps, ranging from -2.8% to -1.3%, while specificity gaps remain within 1.1%–2.2%, suggesting comparatively limited sensitivity to this cue family in the grouped evaluation.

Partial evidence exhibits the strongest and most consistent recall asymmetry. Across all four models, recall is substantially higher for cue-present target-positive examples, with gaps ranging from -33.0% to -20.9%. At the same time, specificity is lower for cue-present negatives, producing gaps of 10.3%–17.7%. Thus, partial evidence is associated with a particularly pronounced dependence on the availability of visible target evidence, affecting both positive recognition and negative discrimination. These results show why aggregate balanced accuracy alone is insufficient to characterize the observed failures and motivate cue-specific diagnosis before remediation.

A.2 Validation and Uncertainty

A.2.1 Cue-level Audit

Three annotators independently judged every visual cue candidate while blinded to the automatic filtering decision and helper confidence. A cue passes a criterion when at least two annotators agree. Table 4 reports macro means across targets. The retained library passes all four criteria under this majority rule. Mean pairwise annotator agreement is 98.4% for visual validity, 94.3% for target relation, 94.9% for non-ambiguity, and 93.9% for family assignment.

A.2.2 Cue–image Audit

The cue-level audit does not establish that each cue is correctly present or absent in every selected image. We therefore sampled 180 unique cue–image items, stratified over the three visual families, four (Y_T, C) cells, and 15 images per family–cell. Three blinded annotators judged target presence and the presence of at least one listed cue. Table 5 compares the automatic group construction with human-majority labels.

A.2.3 Helper-model Sensitivity

Table 6 changes the helper model while holding targets, generation prompts, and filtering criteria fixed. Each helper generates a separate candidate pool and all helpers filter every pool. Cross-filter agreement and Jaccard overlap are therefore evidence about upstream library construction, not a rerun of downstream diagnosis or remediation with a new library.

A.3 Additional Remediation Results

A.3.1 Generic debiasing baselines

We adapt three taxonomy-agnostic prompting strategies from MM-SpuBench (Ye et al., 2026a) as baselines: describe-then-decide, explain-then-decide, and generic debiasing. The prompts used in our experiments are listed below. Results are reported in Table 8.

Table 3: Full grouped visual diagnosis (%). All gaps are $Metric_{C=0} - Metric_{C=1}$ within the corresponding target label.

Cue family	Model	BAcc	$\text{Recall}_{C=0}$	$\text{Recall}_{C=1}$	Recall_{gap}	$\text{Spec}_{C=0}$	$\text{Spec}_{C=1}$	Spec_{gap}
Competitor	Qwen2.5	90.1	87.9	78.1	9.8	99.6	94.7	5.0
Competitor	Qwen3	89.9	93.0	79.5	13.5	98.9	88.3	10.6
Competitor	Gemma-4	88.9	93.1	80.7	12.4	98.0	83.7	14.3
Competitor	GLM-4.6V	93.2	94.8	90.1	4.7	98.6	89.3	9.3
Context	Qwen2.5	92.4	84.7	86.9	-2.1	99.5	98.5	1.1
Context	Qwen3	94.9	90.5	93.3	-2.8	98.9	96.9	2.0
Context	Gemma-4	94.6	91.1	93.5	-2.4	98.1	96.0	2.1
Context	GLM-4.6V	95.6	93.5	94.8	-1.3	98.1	95.9	2.2
Partial evidence	Qwen2.5	84.0	59.0	92.0	-33.0	97.5	87.2	10.3
Partial evidence	Qwen3	85.7	74.9	97.2	-22.3	94.2	76.5	17.7
Partial evidence	Gemma-4	86.0	74.6	95.5	-20.9	92.3	81.5	10.9
Partial evidence	GLM-4.6V	86.7	75.0	98.3	-23.2	94.2	79.2	15.1

Table 4: Majority-vote cue-level audit pass rates (%).

Audited set	Cues	Visual	Target relation	Non-ambiguous	Family correct
Retained library	146	100.0	100.0	100.0	100.0
All visual candidates	209	100.0	99.6	100.0	100.0

1. **Describe-then-decide.** "First, briefly describe the image, focusing on details relevant to whether `{target}` is present. Be mindful of spurious attributes. Then give your final decision on a separate line using exactly: Answer: Yes or Answer: No."

2. **Explain-then-decide.** "Determine whether `{target}` is present in the image. Briefly explain which visible information supports or refutes the presence of `{target}`. Then give your final decision on a separate line using exactly: Answer: Yes or Answer: No."

3. **Generic debiasing.** "Is there a `{target}` in the image? Ensure that your decision is not affected by spurious attributes. Answer using exactly Yes or No."

A.3.2 Prompt-strength Ablation

We vary the strength of cue-matched interventions to examine the trade-off between correcting cue sensitivity and preserving recognition. Soft interventions provide brief reminders about the diagnosed cue, whereas stronger interventions impose increasingly explicit constraints on how the cue should be treated during recognition. This ablation tests whether stronger constraints consistently improve performance on the affected groups or instead over-constrain the model and reduce recall.

Results for the Qwen-series models are reported in Table 9. The results show that increasing intervention strength does not yield monotonic improvements. Medium-strength prompts generally provide the largest gains, while the hard prompt can over-intervene and substantially degrade performance. For example, hard competitor remediation reduces Qwen2.5 balanced accuracy from 90.1% to 85.4%, whereas the medium intervention increases it to 91.1%. Similar patterns appear for context and partial-evidence cues, where the medium intervention achieves the highest or near-highest balanced accuracy across the evaluated models. These results suggest that cue diagnosis should be paired with appropriately calibrated remediation rather than simply imposing stronger constraints on the model.

Table 5: Cue–image audit results (%). False inclusion is automatic $C = 1$ but human-majority $C = 0$; false exclusion is the reverse.

Family	Target agreement	Cue agreement	False inclusion	False exclusion	Exact (Y_T, C)
Context	98.3	91.7	10.0	6.7	90.0
Partial evidence	95.0	95.0	3.3	6.7	91.7
Competitor	98.3	95.0	6.7	3.3	93.3
Overall	97.2	93.9	6.7	5.6	91.7

Table 6: Helper sensitivity under fixed prompts. Self-filter is the accepted percentage. Mean cross-filter agreement is 86.9%–88.4% and accepted-set Jaccard overlap is 84.6%–86.1%.

Generator	Candidates	Completed	Repairs	Self-filter
Qwen3.5-9B	926	96/96	1	83.3
OLMo-3.1-32B	917	96/96	19	83.5
OLMo-3-7B	876	96/96	13	70.4
GPT-OSS-20B	916	96/96	0	78.1

Table 7: Cue-matched prompt remediation results.

Cue family	Model	ΔBAcc (mean $\pm$ std.)
Competitor	Qwen2.5	$+0.92 \pm 0.21$
Competitor	Qwen3	$+1.75 \pm 0.30$
Competitor	Gemma-4	$+2.65 \pm 0.24$
Competitor	GLM-4.6V	$+1.35 \pm 0.33$
Context	Qwen2.5	$+0.73 \pm 0.11$
Context	Qwen3	$+1.40 \pm 0.15$
Context	Gemma-4	$+1.70 \pm 0.27$
Context	GLM-4.6V	$+1.05 \pm 0.32$
Partial evidence	Qwen2.5	$+0.73 \pm 0.25$
Partial evidence	Qwen3	$+1.88 \pm 0.43$
Partial evidence	Gemma-4	$+2.25 \pm 0.38$
Partial evidence	GLM-4.6V	$+2.25 \pm 0.29$

B Prompt Templates

This section provides the prompt templates used by CuE. The prompts are divided into two stages. The offline stage uses a helper LLM to construct and validate target-conditioned hypothesis libraries. The online stage queries the frozen MLLM as a verifier. All online verifier prompts are designed to elicit constrained binary responses so that the response parser can convert responses into comparable yes/no decisions.

B.1 Notation

We use `{target}` for the queried target concept, `{cue}` for a candidate nuisance cue, `{competitor}` for a nearby competing concept, and `{image}` for the input image. Unless otherwise specified, the target MLLM is instructed to answer with `Yes` or `No`.

B.2 Offline Helper-LLM Prompts

Target normalization. This prompt canonicalizes an arbitrary user query into a target concept and removes task-irrelevant wording.

> You are helping build a robustness enhancer for open-vocabulary visual recognition.
>
> Given the user query:
>
> `{query}`
>
> Extract the visual target concept being queried. Return a JSON object with:
>
> - `target`: the canonical target name;
> - `aliases`: common synonyms or paraphrases;
> - `not_targets`: nearby concepts that should not be counted as the target;
> - `query_type`: one of `object`, `attribute`, `part`, `action`, `scene`, or `other`.
>
> Only include visually checkable concepts. Do not include assumptions from language priors.

Target-defining evidence generation. This prompt constructs direct visual evidence hypotheses.

Table 8: Complete remediation comparison on the grouped visual evaluations. Values are percentage-point changes relative to the base query; positive values indicate improvement.

Cue family	Model	Method	ΔBAcc	$\Delta\text{Recall}_{C=0}$	$\Delta\text{Recall}_{C=1}$	$\Delta\text{Spec}_{C=0}$	$\Delta\text{Spec}_{C=1}$
Competitor	Qwen2.5	CuE	+0.92	-0.10	+2.30	-0.10	+1.60
Competitor	Qwen2.5	Describe	+0.40	+3.00	+6.50	-1.30	-6.60
Competitor	Qwen2.5	Explain	+0.05	+1.40	+1.10	-0.70	-1.60
Competitor	Qwen2.5	Generic debiasing	-0.28	-0.60	-0.20	+0.00	-0.30
Competitor	Qwen3	CuE	+1.75	+0.10	+2.70	-0.20	+4.40
Competitor	Qwen3	Describe	-0.13	+2.80	+2.90	-1.50	-4.70
Competitor	Qwen3	Explain	+0.00	+2.30	+3.20	-1.50	-4.00
Competitor	Qwen3	Generic debiasing	-0.15	-0.10	+0.00	+0.00	-0.50
Competitor	Gemma-4	CuE	+2.65	-0.80	+3.40	+0.50	+7.50
Competitor	Gemma-4	Describe	-0.93	-0.70	-4.20	-0.30	+1.50
Competitor	Gemma-4	Explain	-2.12	-0.60	-10.00	-0.30	+2.40
Competitor	Gemma-4	Generic debiasing	-0.28	+1.00	+2.00	-0.50	-3.60
Competitor	GLM-4.6V	CuE	+1.35	+0.80	+0.30	+0.40	+3.90
Competitor	GLM-4.6V	Describe	-0.35	-2.50	-5.20	+0.60	+5.70
Competitor	GLM-4.6V	Explain	+2.20	-0.20	+5.30	+0.30	+3.40
Competitor	GLM-4.6V	Generic debiasing	+0.25	-0.50	-0.60	+0.20	+1.90
Context	Qwen2.5	CuE	+0.73	+2.20	+1.20	-0.20	-0.30
Context	Qwen2.5	Describe	+1.50	+5.60	+4.40	-1.50	-2.50
Context	Qwen2.5	Explain	+0.55	+2.70	+2.00	-0.70	-1.80
Context	Qwen2.5	Generic debiasing	-0.07	-0.30	+0.00	+0.00	+0.00
Context	Qwen3	CuE	+1.40	+3.20	+2.80	-0.10	-0.30
Context	Qwen3	Describe	-0.25	+3.00	+2.00	-2.00	-4.00
Context	Qwen3	Explain	+0.10	+3.10	+1.90	-1.40	-3.20
Context	Qwen3	Generic debiasing	+0.00	+0.00	+0.00	+0.00	+0.00
Context	Gemma-4	CuE	+1.70	+3.30	+2.10	+0.30	+1.10
Context	Gemma-4	Describe	-1.00	-3.00	-2.40	+0.80	+0.60
Context	Gemma-4	Explain	-0.27	-1.00	-0.80	+0.40	+0.30
Context	Gemma-4	Generic debiasing	-0.22	-0.70	-0.60	+0.20	+0.20
Context	GLM-4.6V	CuE	+1.05	+2.50	+1.90	+0.00	-0.20
Context	GLM-4.6V	Describe	-1.10	-1.70	-2.10	+0.00	-0.60
Context	GLM-4.6V	Explain	-0.78	-1.20	-1.50	+0.00	-0.40
Context	GLM-4.6V	Generic debiasing	+0.20	+1.30	+0.80	-0.30	-1.00
Partial evidence	Qwen2.5	CuE	+0.73	-0.60	+0.00	+0.30	+3.20
Partial evidence	Qwen2.5	Describe	-2.00	+7.30	+2.00	-4.50	-12.80
Partial evidence	Qwen2.5	Explain	-1.23	+1.30	+0.30	-1.60	-4.90
Partial evidence	Qwen2.5	Generic debiasing	-0.15	-1.00	+0.10	+0.10	+0.20
Partial evidence	Qwen3	CuE	+1.88	+2.00	+0.30	-0.30	+5.50
Partial evidence	Qwen3	Describe	-1.08	+9.20	-0.10	-5.40	-8.00
Partial evidence	Qwen3	Explain	-0.68	+9.20	+0.60	-4.50	-8.00
Partial evidence	Qwen3	Generic debiasing	-0.17	-0.30	+0.10	-0.20	-0.30
Partial evidence	Gemma-4	CuE	+2.25	+3.50	+0.00	+1.10	+4.40
Partial evidence	Gemma-4	Describe	-1.05	-9.00	-1.80	+3.00	+3.60
Partial evidence	Gemma-4	Explain	-0.55	-2.80	-0.20	+0.60	+0.20
Partial evidence	Gemma-4	Generic debiasing	-0.62	-1.60	-0.40	-0.70	+0.20
Partial evidence	GLM-4.6V	CuE	+2.25	+5.70	-0.90	+1.30	+2.90
Partial evidence	GLM-4.6V	Describe	-0.27	+0.60	-2.40	-1.50	+2.20
Partial evidence	GLM-4.6V	Explain	+0.50	-0.20	-1.60	+0.20	+3.60
Partial evidence	GLM-4.6V	Generic debiasing	-0.22	+4.20	+0.30	-1.80	-3.60

Table 9: Balanced accuracy (%) under varying cue-matched prompt strengths.

Cue family	Model	Base	Soft	Medium	Hard
Competitor	Qwen2.5	90.1	90.7	91.1	85.4
Competitor	Qwen3	89.9	91.0	91.6	89.3
Context	Qwen2.5	92.4	92.3	93.1	91.6
Context	Qwen3	94.9	95.1	96.3	94.6
Partial evidence	Qwen2.5	84.0	84.6	84.7	83.9
Partial evidence	Qwen3	85.7	87.9	87.6	85.5

List visual evidence that can directly verify the presence of `{target}`. Each item should describe intrinsic object-defining evidence, not a single ambiguous part, background, accessory, or co-occurring object. Return exactly `N` items. All items together must define the `{target}`. Each item should be:

- visually observable;
- target-defining rather than merely contextual;
- robust across viewpoint, style, and background;
- short and concrete.

> Do not include common backgrounds, co-occurring objects, locations, or social stereotypes.
> Return JSON with key `direct_evidence`. Each entry should contain: `name`, `description`, and `why_target_defining`.

Scene-context cue generation. This prompt proposes backgrounds or environments that may spuriously support the target.

> For the target concept `{target}`, list scene-context cues that commonly appear with it but are not part of it.
> A scene-context cue is an environment, habitat, background, location, or situational context that may make a model more likely to predict the target even when the target is absent.
> Return exactly `N` items. Each item must be visually observable and must be able to appear without the target.
> Return JSON with key `scene_context_cues`. Each entry should contain: `name`, `description`, and `why_spurious`.

Partial-evidence cue generation. This prompt proposes parts or attributes that may be mistaken for sufficient evidence.

> For the target concept `{target}`, list partial-evidence cues that may suggest the target but do not by themselves prove that the target is present.
> A partial-evidence cue can be a part, texture, attribute, accessory, material, or correlated component. It should be visually observable and should sometimes appear without the target.
> Return exactly `N` items. Do not include cues that are always inseparable from the target.
> Return JSON with key `partial_evidence_cues`. Each entry should contain: `name`, `description`, and `why_insufficient`.

Competitor generation. This prompt proposes nearby concepts that may substitute for the target.

> For the target concept `{target}`, list visually, semantically, or functionally similar concepts that an MLLM may confuse with the target.
> A competitor should be a plausible alternative explanation for an image. It should not be a mere background cue. It should be possible for a human to distinguish the competitor from the target using visual evidence.
> Return exactly `N` items.
> Return JSON with key `competitors`. Each entry should contain: `name`, `shared_evidence`, `distinguishing_evidence`, and `why_confusable`.

Language-side cue generation. This prompt constructs query rewrites and language-prior probes.

> For the target concept `{target}`, generate language-side robustness checks.
> Return JSON with three keys:
>
> - `query_rewrites`: semantically equivalent yes/no questions about whether the target is visible;
> - `language_prior_probes`: prompts that test whether the target may be predicted from common co-occurrences or textual plausibility rather than visual evidence;
> - `text_interference_checks`: prompts that distinguish direct visual evidence from target mentions in OCR, captions, surrounding text, or user-provided descriptions.
>
> All query rewrites must preserve the same visual target. Avoid adding new assumptions, contexts, or attributes.

B.3 Offline Filtering Prompts

Visual detectability filter.

> Determine whether `{cue}` is visually detectable in an image.
>
> A cue is visually detectable if it has a physical or visible manifestation and a human can usually identify which pixels correspond to it. Abstract concepts, lighting effects without clear boundaries, emotions, and non-visual associations should be marked `No`.
>
> Answer with `Yes` or `No` only.

Target-part filter.

> Is `{cue}` physically part of `{target}`?
>
> Answer `Yes` if `{cue}` is a component, subpart, material, or inseparable attribute of `{target}`. Answer `No` if it is only commonly nearby, contextually related, or visually similar.
>
> Answer with `Yes` or `No` only.

Cue-without-target filter.

> Can `{cue}` appear in an image without `{target}` being present?
>
> Answer with `Yes` or `No` only.

Target-without-cue filter.

> Can `{target}` appear in an image without `{cue}` being present?
>
> Answer with `Yes` or `No` only.

Synonym filter.

> Do `{cue}` and `{target}` generally refer to the same object or concept, or is one merely a synonym or instance of the other?
>
> This is not asking whether they are related or often appear together.
>
> Answer with `Yes` or `No` only.

Competitor validity filter.

> Could `{competitor}` be visually or functionally confused with `{target}` in an image, while still being a distinct concept?
>
> Answer `Yes` only if both conditions hold:
>
> - the concepts are distinct;
> - they share enough visual, functional, or contextual evidence that an MLLM might substitute one for the other.
>
> Answer with `Yes` or `No` only.

Query-equivalence filter.

> Do the following two questions ask for the same visual target, without adding new context, attributes, or assumptions?
>
> Question 1: `{query_1}`
>
> Question 2: `{query_2}`
>
> Answer with `Yes` or `No` only.

B.4 Online MLLM Remediation Prompts

CuE uses the frozen target MLLM itself for both base recognition and cue-conditioned remediation. All online prompts are rendered as binary yes/no questions. Unless a prompt already contains the instruction, CuE appends: `Answer with Yes or No only.` The system prompt is: `Answer the user's question with Yes or No only.` The base recognition prompt is:

> Is there `{a/an target}` in the image?

Direct-evidence remediation. Direct-evidence remediation asks the model to base the decision on target-defining visual evidence rather than contextual or partial cues. CuE evaluates one generic direct-evidence family and three helper-conditioned strengths. The generic family uses no helper-generated cue:

> Ignoring background, context, and common co-occurring objects, is there direct visual evidence of `{target}` in this image?

The helper-soft variant provides target-defining evidence as a helpful clue:

> A helpful clue for recognizing `{target}` is: `{direct_evidence}`. Is there `{a/an target}` in the image?

The helper-medium variant explicitly prioritizes object structure over surrounding context:

> Answer mainly from the object's visible structure rather than surrounding context. A helpful clue for `{target}` is: `{direct_evidence}`. Is there `{a/an target}` in the image?

The helper-hard variant corresponds to the checklist-style direct-evidence prompt:

> For `{target}`, check only the following direct visual evidence: `{direct_evidence}`. Is there enough evidence to verify that `{target}` is present?

Scene-context remediation. Scene-context remediation discourages context-only recognition. CuE constructs soft, medium, and hard variants from accepted scene-context cues.

Soft:

> Helpful scene context for `{target}` includes `{context_cues}`. Is there `{a/an target}` in the image?

Medium:

> Scene context such as `{context_cues}` can support recognizing `{target}`, but it is not sufficient by itself. Is there `{a/an target}` in the image?

Hard:

> Is there `{a/an target}` in the image? Do not answer Yes from scene context alone. Even if scene context such as `{context_cues}` is present, answer Yes only if `{a/an target}` is actually visible in the image.

Partial-evidence remediation. Partial-evidence remediation warns that local fragments or associated objects are not sufficient for target presence. CuE evaluates a generic caution prompt and helper-conditioned soft, medium, and hard variants.

Generic:

> Do not answer Yes from partial cues, fragments, or nearby suggesting objects alone. Is there `{a/an target}` in the image?

Soft:

> Helpful but non-decisive cues for `{target}` include `{partial_cues}`. Is there `{a/an target}` in the image?

Medium:

> Cues such as `{partial_cues}` can suggest `{target}`, but they are insufficient on their own. Is there `{a/an target}` in the image?

Hard:

> Is there `{a/an target}` in the image? Do not answer Yes from partial cues alone. Even if `{partial_cues}` is visible, answer Yes only if `{a/an target}` itself is actually visible in the image.

Competitor-aware remediation. Competitor-aware remediation asks the model to distinguish the queried target from a resolved competing concept. CuE evaluates soft, medium, and hard competitor-aware target prompts.

Soft:

> Be careful not to confuse `{target}` with `{competitor}`. Is there `{a/an target}` in the image?

Medium:

> `{competitor}` may be a distractor. Is there `{a/an target}` in the image?

Hard:

> Is there `{a/an target}` in the image? Do not answer Yes merely because `{a/an competitor}` is present. Answer Yes only if `{target}`-specific visual evidence is directly visible, even if `{competitor}` is also present. Key distinction: `{distinguishing_evidence}`.

Language-side prompts. For query-form evaluation, CuE uses accepted helper-generated query rewrites directly after yes/no rendering. For language-prior probes, helper-generated prompts are either used directly if they are already visual recognition questions, or anchored with the base visual question:

> `{language_prior_probe}` Based only on the visible image, is there `{a/an target}` in the image?

For text-interference evaluation, CuE uses accepted helper-generated prompts directly, grouped by subtype such as irrelevant fact, contradictory fact, user judgment, competitor suggestion, and instructional bias. These probes test whether misleading or distracting textual conditions change the model's image-grounded recognition decision.

C Limitations and Future Work

Several limitations remain. First, the proposed cue-family taxonomy is not exhaustive. This paper focuses on visual and linguistic cue families that are central to zero-shot MLLM recognition, including scene context, partial evidence, competitors, query form, language priors, and text interference. Other robustness factors, such as viewpoint, crop, blur, resolution, lighting, rendering style, and synthetic image artifacts, are not included in the present evaluation and should be studied in future work.

Second, our evaluation uses a deliberately stratified 12-target COCO panel and four MLLMs from three model families. This setting provides controlled cross-category evidence, but it does not cover the full diversity of open-vocabulary targets, datasets, domains, and model families. Future work

should extend the evaluation to larger, systematically sampled target vocabularies and additional datasets.

Third, the cue groups use naturally occurring images rather than matching object size, occlusion, viewpoint, salience, and scene complexity across cue conditions. The cue–image audit quantifies residual group-construction noise, but it does not remove these potential recognition-difficulty confounds. Matched-subset evaluation or explicit difficulty annotation in future work would strengthen the causal interpretation of the visual gaps.

Finally, the current remediation strategy is cue-family-wise. We evaluate prompt interventions separately for each diagnosed cue family, which preserves interpretability but does not yet provide an orchestration mechanism for images affected by multiple cue families simultaneously. A natural next step is to compose cue-specific prompts into a multi-branch solver that selects or combines interventions while retaining family-level diagnostic traceability.

D Computational Environment

All experiments were run on a Linux server with Ubuntu 22.04.5 LTS. The server has 2× Intel Xeon Gold 6348 CPUs at 2.60 GHz, with 28 cores per socket, 2-way SMT, and 112 hardware threads in total. It has 1.5 TiB of RAM and 8 GiB of swap. GPU inference was performed on 4× NVIDIA L40 GPUs, each with 48 GiB of VRAM, using NVIDIA driver 555.42.02 and CUDA 12.5.