CoAMA: A Hierarchical Benchmark for Diagnosing Cross-Object Attribute Misbinding in Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large vision-language models (LVLMs) are increasingly used for visual understanding and decision-making, yet they may still produce plausible but incorrect visual interpretations—a phenomenon known as hallucination. In multi-object scenes, such failures arise when an attribute belonging to one object is incorrectly assigned to another, undermining reliable scene understanding and downstream reasoning. However, existing hallucination benchmarks mainly evaluate object existence, class recognition, or attribute presence, while failing to evaluate whether perceived attributes are correctly bound to their corresponding object instances. To address this gap, we introduce **C**ross-**o**bject **A**ttribute **M**isbinding **A**ssessment (CoAMA), a hierarchical diagnostic benchmark for evaluating instance-level attribute binding through three progressive levels: multi-object instance recognition, target-object attribute recognition, and cross-object attribute binding verification. Its core design employs paired in-image attribute-swap probes, where each negative probe transfers a ground-truth attribute from one marked object to another, directly testing whether the model preserves the instance-level ownership of attributes that are visually present in the same scene. CoAMA contains 787 multi-object images, 3,935 marked instances, and 19,460 scorable attributes, with standardized annotation and scoring protocols enabling consistent and reproducible evaluation. We evaluate representative LVLMs and find that even the best-performing model, Qwen3.7-Plus, achieves only 67.69% PairAcc under L3-Oracle-Direct, while most models exhibit higher false-acceptance rates for attributes belonging to other marked objects than for attributes absent from the marked-object set. These findings reveal cross-object attribute misbinding as a critical reliability challenge and provide a controlled foundation for improving LVLM reliability in complex visual environments.

1 Introduction

Large vision-language models (LVLMs) have made substantial progress in understanding and reasoning about complex visual content (Liu et al., 2023). As these models are increasingly used in embodied agents (Driess et al., 2023), autonomous systems (Sima et al., 2024), and robotic manipulation (Zitkovich et al., 2023), reliable understanding of multi-object scenes becomes essential. Reliable multi-object understanding requires models not only to recognize objects and attributes, but also to preserve the correspondence between each attribute and the specific object instance it describes.

Despite strong recognition performance, LVLMs remain vulnerable to *cross-object attribute misbinding*, a failure in which a model assigns an attribute present in the image to the wrong object instance. For example, given an image containing a silver spoon and a green broccoli floret, a model may correctly recognize the spoon and the color green, yet incorrectly answer that the spoon is green. Such errors are difficult to detect because both the queried object and the predicted attribute are visually present, while only their correspondence is incorrect. This failure constitutes a fine-grained form of multimodal hallucination and exposes a gap between recognizing visual elements and organizing them into a structured, instance-level representation (Yuksekgonul et al., 2023; Pham & Schott, 2024).

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.

Most existing hallucination benchmarks assess whether an object, attribute, or relation is present or factually correct, but do not directly determine whether a perceived attribute belongs to the queried object instance. TruthfulQA and HalluLens characterize general factuality and intrinsic or extrinsic hallucination (Lin et al., 2022; Bang et al., 2025), POPE focuses on object-existence judgments (Li et al., 2023), and AMBER extends hallucination evaluation to attributes and relations (Wang et al., 2023). ROPE further studies cross-object interference (Chen et al., 2024), while H-POPE constructs adversarial questions using attributes drawn from other objects in the same image (Pham & Schott, 2024). These benchmarks still expose a critical gap between recognition capability and structured visual understanding. Beyond identifying what is present in an image, a model must also determine which attributes are associated with which object instances.

These limitations leave instance-level attribute ownership insufficiently characterized in multi-object scenes. To address this gap, we introduce **C**ross-**o**bject **A**ttribute **M**isbinding **A**ssessment (CoAMA), a hierarchical diagnostic benchmark for multi-object scenes. It decomposes visual understanding into three progressive levels: (i) multi-object instance recognition, (ii) target-object attribute recognition, and (iii) cross-object attribute binding verification. CoAMA contains 787 multi-object images, 3,935 marked instances, and 19,460 scorable attributes annotated under unified criteria. By organizing recognition and binding into distinct evaluation tasks, this hierarchy isolates instance-level attribute binding as a distinct evaluation target and supports targeted diagnosis of binding performance and upstream error propagation. At Level 3, CoAMA constructs paired in-image attribute-swap probes. A positive proposition queries a ground-truth attribute of the target, whereas its paired swapped negative assigns to the same target a ground-truth attribute value drawn from another object, with the two probes sharing the same image, target object, and attribute type, differing only in the queried value. Figure 1 summarizes the hierarchical evaluation pipeline and paired attribute-swap construction.

To disentangle cross-object interference from attribute-specific effects, CoAMA further compares swapped attributes that are present on other marked objects with attribute values that are absent from all marked objects. Additional variants examine target-reference conditions, upstream error propagation, and visual-evidence availability.

We evaluate eight representative LVLMs using CoAMA. The results reveal a persistent recognition–binding gap: models that perform comparatively well in object-instance and target-attribute recognition still make substantial errors in cross-object attribute binding verification. For most evaluated models, swapped-present attributes from other marked objects produce higher false-acceptance rates than absent attributes.

Our main contributions are as follows:

- We establish cross-object attribute misbinding as a distinct reliability challenge in structured scene understanding, showing that correctly recognizing objects and their attributes does not necessarily ensure reliable attribute-to-instance correspondence.
- We introduce CoAMA, a hierarchical diagnostic benchmark that unifies multi-object instance recognition, target-object attribute recognition, and cross-object attribute binding verification. CoAMA operationalizes attribute ownership as a controlled, target-specific verification task and supports fine-grained analysis across diverse object, attribute, and scene conditions.
- Our evaluation shows that cross-object attribute binding remains unreliable across the evaluated models. Controlled comparisons further reveal that attributes grounded in other marked objects produce higher false-acceptance rates than attributes absent from the marked-object set, identifying a systematic source of cross-object interference in LVLMs.

2 Related Work

2.1 LVLM Hallucination Evaluation

Hallucination evaluation for large vision-language models examines whether model outputs remain grounded in the visual input. Early work primarily focused on object hallucination. CHAIR measures object hallucination in image captioning using image annotations (Rohrbach et al., 2018), whereas POPE formulates evaluation as binary object-existence questions (Li et al., 2023). ALOHa

extends caption-level evaluation to open-vocabulary settings (Petryk et al., 2024). For free-form responses, FaithScore evaluates visual faithfulness, while MMHal-Bench and HallusionBench assess hallucinations in open-ended image reasoning (Jing et al., 2024; Sun et al., 2024; Guan et al., 2024).

Recent benchmarks move beyond isolated object presence. ROPE uses visual referring prompts and controlled outputs to test concurrent recognition of multiple marked instances, showing that multi-object probing is more fragile than independent single-object queries (Chen et al., 2024). R-Bench evaluates relationship hallucinations through image-level and instance-level questions (Wu et al., 2024), while Reefknot characterizes relation hallucinations from perceptive and cognitive perspectives (Zheng et al., 2025). These works broaden hallucination evaluation from object existence to multi-object and relational understanding.

2.2 Fine-Grained Visual Attribute Evaluation

Fine-grained benchmarks examine attributes and relations beyond object existence. AMBER evaluates generative and discriminative hallucinations across existence, attribute, and relation dimensions (Wang et al., 2023). M-HalDetect provides fine-grained annotations for unfaithful descriptions, while FIHA generates questions about objects, attributes, relations, and their dependencies (Gunjal et al., 2024; Yan et al., 2025).

Related diagnostic benchmarks assess whether vision-language models preserve distinctions among objects and their properties. VL-CheckList, VALSE, and Winoground probe object, attribute, relation, and compositional understanding under controlled settings (Zhao et al., 2022; Parcalabescu et al., 2022; Thrush et al., 2022). ARO exposes failures in linking objects to their attributes (Yuksekgonul et al., 2023), whereas SugarCrepe constructs fluent hard negatives with adversarial refinement to reduce linguistic shortcuts (Hsieh et al., 2023).

H-POPE extends polling from object existence to attributes and introduces an image-based adversarial setting in which a negative attribute may appear elsewhere in the image (Pham & Schott, 2024). However, it references the target by class and does not preserve an explicitly marked target–source pair, leaving the source instance and attribute-transfer direction unspecified. CoAMA addresses this gap by evaluating cross-object attribute misbinding over marked instances through Multi-Object Instance Recognition, Target-Object Attribute Recognition, and Cross-Object Attribute Binding Verification.

3 The CoAMA Benchmark

CoAMA is a hierarchical benchmark for evaluating instance-level attribute binding in multi-object scenes. It organizes the evaluation into three levels, namely **Multi-Object Instance Recognition**, **Target-Object Attribute Recognition**, and **Cross-Object Attribute Binding Verification**. Levels 1 and 2 establish whether a model can identify the marked instances and recognize the explicit attributes of a specified target. Level 3 evaluates attribute ownership through **paired in-image attribute-swap probes**, in which an attribute that is visually present but belongs to another marked object is queried against the target (Figure 1).

3.1 Task Definition

Let I denote an image containing $K = 5$ marked object instances, indexed by $i \in \{1, \dots, K\}$. The visual prompt set $P = \{p_i\}_{i=1}^{K}$ contains one numbered red bounding box p_i for each instance. Instance i has ground-truth class c_i and canonical attribute values $\{a_i^t\}_{t \in \mathcal{T}}$, where t indexes the attribute type and $\mathcal{T}$ comprises color, material, shape, physical state, and pose.

For a target i, a source $j \neq i$, and an attribute type t, a candidate pair requires scorable values with $a_i^t \neq a_j^t$, following the annotation rules in Section 3.2. Each query tuple encodes the target index, attribute type, and queried value. The positive query q^+ and swapped-negative query q^- have reference labels y^+ and y^-, respectively,

$$q_{i,t}^{+} = (i, t, a_i^t), \quad y^+ = \text{Yes}; \qquad q_{i,j,t}^{-} = (i, t, a_j^t), \quad y^- = \text{No}.$$

The questions share the image, target, and attribute type, changing only the queried value. The queried value is therefore grounded in the image, but its annotated owner differs from the queried

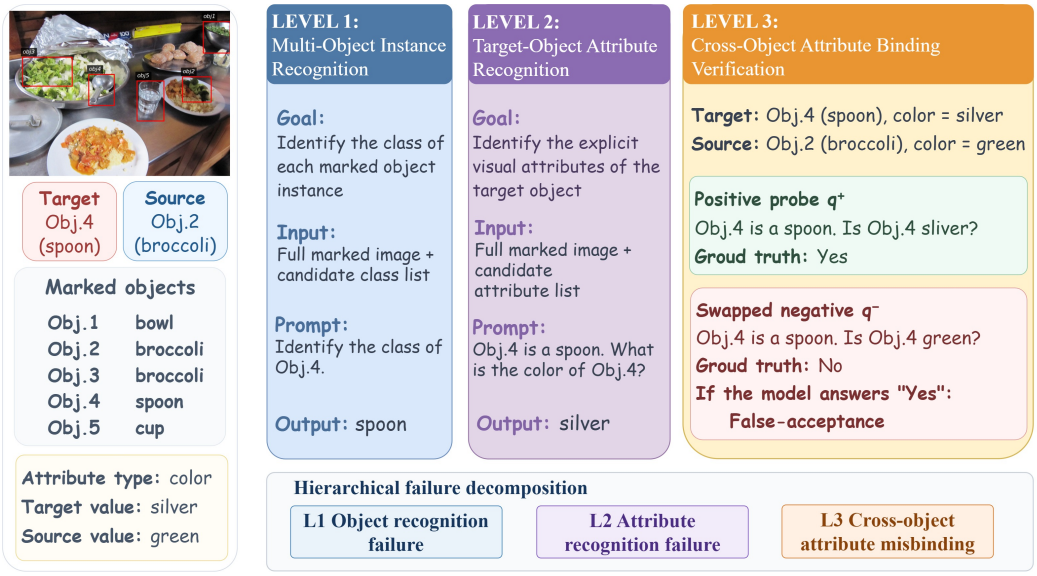


Figure 1: Overview of CoAMA on five marked instances. Level 1 jointly predicts all five classes (the panel illustrates one assignment); Level 2 predicts a target attribute; Level 3 verifies paired target and source-derived values.

target. This construction distinguishes source-derived negatives from those based on attributes absent from the marked-object set. Cross-object attribute misbinding occurs when the model accepts this transferred value as an attribute of the target.

3.2 Dataset Construction and Quality Validation

CoAMA uses the ROPE multi-object image pool (Chen et al., 2024), drawn from COCO and ADE20K (Lin et al., 2014; Zhou et al., 2017). Its 787 images contain 3,935 fixed, numbered instances and 19,460 scorable attributes. The benchmark contains five scene-composition splits. Homogeneous contains five marked objects from the same class, whereas Heterogeneous contains five different object classes. AAAAB and BAAAA each contain four same-class objects and one different-class object, with the different object appearing at different positions in the query order. Mixed retains naturally occurring class compositions.

Each image contains five objects marked with red bounding boxes and indexed as Obj.1 through Obj.5. The visual mark and its object index define a unique instance anchor that is preserved throughout annotation, prompting, response parsing, and scoring.

Each marked instance is annotated with one value for every attribute type in $\mathcal{T}$. Annotators determine the dominant visible color, exterior material, shape, physical state, and pose from evidence within the marked region. Attributes obscured by occlusion, poor image quality, ambiguity, or insufficient evidence are labeled unknown and excluded from attribute scoring and Level 3 probe construction. Material refers to the visible exterior substance of the marked object, whereas physical state describes the physical condition of the object itself. Container contents and surrounding environmental states are not transferred to the marked instance. The labels mixed and irregular are retained for visibly composite colors and shapes. Two annotators independently labeled all 19,675 instance–attribute entries, obtaining macro-averaged Cohen's $\kappa = 0.952$ (Cohen, 1960); disagreements, including those involving unknown labels, were resolved through adjudication. The full vocabulary and annotation rules are provided in Appendix A.2 and Table S2.

3.3 Hierarchical Evaluation Protocol

Level 1: Multi-Object Instance Recognition. The model receives the full image with all five red boxes and a candidate class list, and predicts the classes of the marked objects in a single response.

Table 1: Diagnostic variants of Level 3.

Group	Variants	Controlled factor
Target reference	L3-Oracle-Spatial; L3-Oracle-ObjOnly	How the same target instance is specified
Error propagation	L3-L2GT-Direct; L3-End2End-Direct	Effect of upstream attribute and hierarchical predictions
Visual evidence	L3-TargetCrop; L3-SourceMasked	Availability of target and source evidence
Negative construction	L3-PresentVsAbsent	Comparison of the two negative-probe types

The response contains five comma-separated object–class assignments in index order, from Obj.1 to Obj.5. Level 1 measures concurrent recognition of the five marked instances. Its predictions are reused in the prediction-conditioned Level 3 setting.

Level 2: Target-Object Attribute Recognition. The model receives the full marked image, the target object index and ground-truth class, and a queried attribute type, and predicts the corresponding attribute value from the predefined taxonomy. Each scorable attribute is queried independently under Single-GT. Level 2 evaluates whether the model can recognize the explicit visual attributes of the specified target object.

Level 3: Cross-Object Attribute Binding Verification. Level 3 uses paired in-image attribute-swap probes. For each valid ordered target–source–attribute triplet (i, j, t), CoAMA constructs a positive probe from the target value and a swapped negative from the source value, as defined in Section 3.1. The two probes share the same image, target object, and attribute type, and differ only in the queried value. Probe construction enumerates ordered target–source candidates for each scorable attribute type, removes candidates with identical or unknown values, and applies the predefined sampling rules. The primary setting, L3-Oracle-Direct, specifies the target by index and ground-truth class and uses a frozen manifest of 2,535 pairs. Appendix B.2 details the construction and evaluation sets.

Building on this oracle construction, Table 1 summarizes the additional variants and their controlled factors. The target-reference variants retain oracle probe construction while changing how the target is described. The error-propagation variants inject predictions from earlier levels to quantify how upstream errors affect binding verification. The visual-evidence controls modify the image while preserving the original target–source definition. Variant-specific prompts and evaluation records are provided in Appendix B.1 and Appendix B.2.

L3-PresentVsAbsent compares source-derived values with values absent from the five marked-object annotations. For example, suppose Obj.1 is a green broccoli floret and Obj.2 is a silver spoon. The positive probe asks whether Obj.1 is green, and the reference answer is yes. Its paired negative asks whether Obj.1 is silver, and the reference answer is no. Silver is visually present and assigned to Obj.2, whereas the queried target remains Obj.1. An absent probe instead asks whether Obj.1 is red, provided no marked object is annotated red. All three questions retain the target and attribute type, while the queried value comes from the target, the source, or outside the marked-set annotations. Thus Oracle-Direct tests the positive–swap pair, whereas PresentVsAbsent contrasts the two negative branches. Marked-set absence does not imply whole-image absence; construction and comparison details are given in Appendix B.3.

3.4 Response Generation, Scoring, and Metric Aggregation

All prompts impose explicit output constraints to support deterministic parsing. Level 1 responses are parsed into five object-class predictions, Level 2 responses into one normalized attribute label, and Level 3 responses into yes or no. Capitalization, punctuation, and accepted lexical variants are normalized using predefined rules, without an external neural judge. Unparseable responses remain in the scoring denominator and are recorded as incorrect for accuracy and PairAcc. Level 1 reports instance-level accuracy, while Level 2 reports accuracy over scorable attributes under Single-GT.

For a Level 3 pair r, let $z_r^+ = 1$ when the model answers yes to the positive probe and $z_r^- = 1$ when it answers no to the swapped negative; otherwise the corresponding indicator is zero. With R denoting the number of complete scored pairs, the primary metric is paired accuracy,

$$\mathrm{PairAcc} = \frac{1}{R}\sum_{r=1}^{R} z_r^+ z_r^- .$$

A pair is counted as correct only when the model accepts the target's annotated attribute and rejects the transferred source attribute. Complementing the binary polling used in POPE and H-POPE (Li et al., 2023; Pham & Schott, 2024), CoAMA also reports positive acceptance ($\mathrm{TAR}_{\mathrm{pos}}$) and false acceptance for swapped-present and marked-set-absent negatives ($\mathrm{FAR}_{\mathrm{swap}}$ and $\mathrm{FAR}_{\mathrm{absent}}$). For branch $b \in \{\mathrm{pos}, \mathrm{swap}, \mathrm{absent}\}$, let Q_b be its evaluated question set and Y_b the number of parsed yes responses. The acceptance rates are

$$\mathrm{TAR}_{\mathrm{pos}} = \frac{Y_{\mathrm{pos}}}{|Q_{\mathrm{pos}}|}, \qquad \mathrm{FAR}_b = \frac{Y_b}{|Q_b|} \quad (b \in \{\mathrm{swap}, \mathrm{absent}\}).$$

Each denominator includes parse failures, which contribute no yes response. Overall PairAcc is micro-averaged over scored pairs and further decomposed by attribute type and scene-composition split. YesRatio is the proportion of yes responses among all evaluated questions in a setting; parse-failure rate is the proportion of unparseable responses. Additional metric conventions and normalization rules are provided in Appendix B.4 and Appendix B.5.

4 Experiments and Diagnostic Analysis

We report the three task scores and then examine target references, upstream predictions, visual evidence, and negative construction under the settings defined in Section 3.3.

4.1 Experimental Setup

Models and coverage. We evaluate eight LVLMs spanning open-weight and proprietary systems, multiple parameter scales, dense and sparse architectures, and standard and reasoning-oriented inference paradigms. These include Llama-3.2-11B-Vision-Instruct (Llama-3.2-11B) (Meta AI, 2024); Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct, and Qwen3-VL-32B-Thinking (Bai et al., 2025); Qwen3.6-35B-A3B (Qwen Team, 2026); MiniCPM-o 4.5 (Cui et al., 2026); Gemini 3.5 Flash (Google DeepMind, 2026); and Qwen3.7-Plus (Alibaba Cloud, 2026). All complete Level 1, Level 2 Single-GT, and L3-Oracle-Direct. The six open-weight models additionally complete the diagnostic variants in Table 1; findings from these variants concern those six models.

Implementation and statistics. All models are evaluated using the same images, visual markers, task instructions, attribute taxonomy, response constraints, parsers, and scoring rules for the shared primary tasks. Models that expose decoding controls use greedy decoding with zero temperature. L3-Oracle-Direct uses a frozen manifest containing 2,535 probe pairs. Existing uncertainty estimates use 2,000 image-clustered percentile bootstrap replicates (95% intervals). Model configurations and the statistical procedure are given in Appendix C. Parsing and unknown-label exclusions follow Section 3.4.

4.2 Overall Results: Cross-Object Attribute Misbinding Remains Prevalent

Figure 2 summarizes performance on the three evaluation tasks. At Level 1, Gemini 3.5 Flash and Qwen3.7-Plus achieve accuracies of 85.64% and 81.37%, respectively, indicating strong multi-object instance recognition. At Level 2 Single-GT, the same two models reach attribute accuracies of 66.56% and 64.58% when the target class is provided.

Level 3 reports PairAcc for paired binary verification. Its scores are not directly comparable with Level 2 attribute accuracy because the tasks use different response spaces, scoring rules, and evaluation sets. Because PairAcc requires both binary judgments in a probe pair to be correct, independent random guessing yields an expected score of 25%. Under L3-Oracle-Direct, PairAcc ranges from 43.04% for Llama-3.2-11B to 67.69% for Qwen3.7-Plus. Notably, Gemini 3.5 Flash leads at both

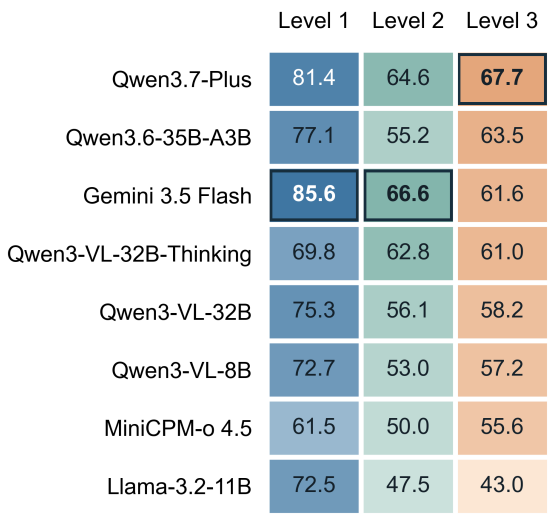


Figure 2: Performance across the three levels of CoAMA. Cells report percentages, and outlined cells mark the best result at each level. Level 3 reports PairAcc under L3-Oracle-Direct.

recognition levels but reaches only 61.58% PairAcc at Level 3. These results indicate that stronger multi-object instance recognition and target-object attribute recognition do not automatically yield reliable cross-object attribute binding verification. Even Qwen3.7-Plus, the strongest model under this condition, answers at least one judgment incorrectly for 32.31% of probe pairs.

These results show that cross-object attribute binding is not reducible to object or attribute recognition alone. A model may identify the marked objects and recover the relevant attributes, yet still fail to preserve the ownership relation between an attribute and its corresponding visual instance. By progressing from what is present to which object owns the perceived evidence, CoAMA establishes instance-level attribute binding as a distinct evaluation target and examines how upstream recognition errors affect binding judgments.

We further examine whether these binding failures vary systematically across probe and scene characteristics. Stratified analyses under L3-Oracle-Direct reveal systematic variation across attributes and scene structure. Pose is the most challenging attribute type, whereas material and physical state are more reliably bound (Table S11). The strongest structural effect arises when the target and source belong to the same class, reducing PairAcc by 16.02 percentage points on average across models (Table S13). Homogeneous scenes are similarly more difficult (Table S12), while target size and target–source distance exhibit weaker and less consistent effects (Table S14).

4.3 Diagnostic Analysis

Table 2 reports the controlled Level 3 results for the six open-weight models and provides the basis for the following diagnostic analyses.

4.3.1 Target-Reference Conditions

The target-reference conditions examine how different forms of target specification affect Level 3 performance. Replacing the direct target specification with a spatial description (L3-Oracle-Spatial) lowers PairAcc for every model, producing a range of 38.30%–57.36%. Removing the target class while retaining only the target index (L3-Oracle-ObjOnly) causes a further decline to 35.98%–50.81%. These consistent degradations suggest that target-class semantics provide useful support for identifying the intended object. Object indices or spatial descriptions alone appear to offer weaker semantic grounding, which may contribute to less stable Level 3 judgments.

Table 2: Controlled Level 3 results. All columns report PairAcc except the final column, which reports FAR for marked-set-absent negatives.

Model	Oracle Direct	Oracle Spatial	Oracle ObjOnly	L2GT Direct	End2End Direct	Target Crop	Source Masked	Absent FAR
Llama-3.2-11B	43.04%	38.30%	37.75%	34.45%	31.53%	42.96%	43.43%	27.95%
Qwen3-VL-8B	57.20%	51.68%	39.49%	42.08%	36.50%	61.34%	57.20%	9.23%
Qwen3-VL-32B	58.19%	52.66%	45.36%	46.52%	44.08%	60.16%	58.82%	10.55%
Qwen3.6-35B-A3B	63.55%	57.36%	50.81%	49.55%	43.81%	64.89%	63.19%	13.01%
Qwen3-VL-32B-Thinking	61.03%	56.69%	48.95%	45.03%	40.53%	61.11%	61.78%	16.80%
MiniCPM-o 4.5	55.58%	48.40%	35.98%	46.01%	35.60%	61.26%	55.78%	14.65%

4.3.2 Upstream Error Propagation

The propagation settings quantify how Level 1 and Level 2 prediction errors compound the Level 3 binding failure. For all six models, PairAcc declines monotonically from L3-Oracle-Direct to L3-L2GT-Direct and then to L3-End2End-Direct.

When predicted Level 2 attributes replace the ground-truth target attributes, PairAcc decreases to 34.45%–49.55%. Introducing predicted Level 1 object identities further reduces PairAcc to 31.53%–44.08%. These consistent declines show that errors in instance recognition and attribute recognition propagate into subsequent binding judgments. Accurate upstream recognition at both levels is therefore necessary for reliable instance–attribute binding, but it is not sufficient to guarantee reliable binding.

4.3.3 Visual-Evidence Availability

We examine whether visual evidence outside the target region contributes to cross-object attribute misbinding. In TargetCrop, only the target object and its immediate visual context are retained, while other objects and most of the surrounding scene are removed. PairAcc improves for five of the six models, increasing from 56.43% under L3-Oracle-Direct to 58.62% on average. The largest gains are observed for MiniCPM-o 4.5, from 55.58% to 61.26%, and Qwen3-VL-8B, from 57.20% to 61.34%. These results show that restricting the visual input to the target region reduces binding errors for most models, suggesting that other objects in the full scene may interfere with attribute binding to the target object.

SourceMasked provides a complementary control by removing only the source object that carries the swapped attribute, while preserving the target object in its original location and retaining the remainder of the scene. Under this condition, the six-model mean PairAcc increases only slightly, from 56.43% to 56.70%. This modest improvement suggests that removing the source evidence may alleviate cross-object attribute misbinding to a limited extent. Since most background and contextual information remains available, source removal alone does not substantially reduce the observed binding errors.

4.3.4 Negative Construction

In L3-Oracle-Direct, each swapped negative replaces the target attribute with an attribute belonging to another marked object in the same image. L3-PresentVsAbsent examines whether an attribute with genuine visual evidence elsewhere in the image is more likely to be falsely accepted than an attribute that is absent from the attribute annotations of all marked objects. Because both conditions consist of negative probes, FAR is the primary metric. We define

$$\Delta_{\mathrm{FAR}} = \mathrm{FAR}_{\mathrm{swap}} - \mathrm{FAR}_{\mathrm{absent}}.$$

As shown in Figure 3, Δ_{FAR} is positive for all six models, ranging from 4.07 to 25.98 percentage points. The L3-PresentVsAbsent construction guarantees only that the queried attribute is absent from the attribute annotations of the current marked-object set; it does not rule out the possibility that the attribute appears in an unmarked region. Despite this conservative definition, every model

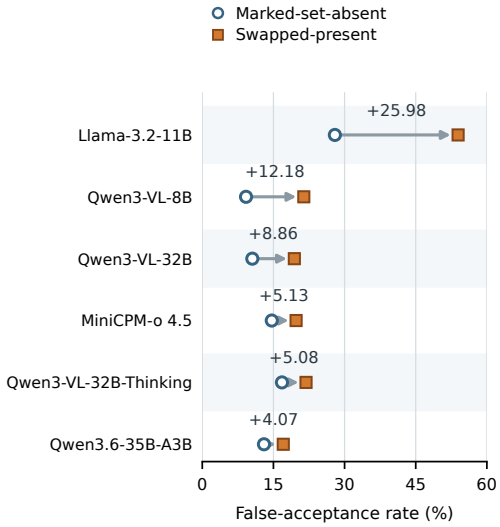


Figure 3: Marked-set-absent and swapped-present false-acceptance rates under L3-PresentVsAbsent. Labels report paired increases in percentage points.

exhibits a higher FAR for swapped-present negatives. This descriptive gap suggests that the errors captured by CoAMA may not be fully explained by a general tendency to accept arbitrary negative attributes. Instead, attributes with genuine visual evidence elsewhere in the image are more likely to be incorrectly bound to the target object, resulting in a higher rate of cross-object attribute misbinding.

The controlled analyses show that Level 3 performance is jointly affected by target specification, upstream recognition quality, and competing visual evidence. The consistently higher false-acceptance rates for swapped-present negatives further indicate that attributes supported elsewhere in the image are particularly likely to be misassigned to the queried target.

5 Conclusion

We introduce CoAMA, a hierarchical diagnostic benchmark for cross-object attribute misbinding in multi-object scenes. Through three progressive levels—multi-object instance recognition, target-object attribute recognition, and cross-object attribute binding verification—CoAMA evaluates upstream recognition and attribute binding and traces the propagation of prediction errors across levels. Across eight representative LVLMs, the strongest model reaches only 67.69% PairAcc under L3-Oracle-Direct, showing that recognizing the relevant objects and attributes does not ensure reliable instance-level binding. This performance gap provides empirical support for treating cross-object attribute binding verification as a necessary evaluation dimension for multi-object vision-language understanding. Overall, reliable multi-object understanding depends not only on recognizing what is present, but also on preserving which instance each piece of visual evidence belongs to. By making this distinction explicit and measurable, CoAMA lays a foundation for developing more trustworthy embodied agents and visual decision-making systems.

AI Use Statement

Generative AI tools were used to assist with language polishing, literature retrieval and discovery, software development, and limited discussion of research methodology and experimental analysis. All research decisions, experimental designs, and interpretations were independently evaluated and finalized by the authors. AI-assisted code was reviewed and tested by the authors, and literature identified with AI assistance was independently verified against the original sources. The authors

reviewed all AI-assisted content and take full responsibility for the final manuscript, claims, code, and research artifacts.

REFERENCES

Alibaba Cloud. Qwen3.7-Plus vision model documentation. https://www.alibabacloud.com/help/en/model-studio/vision-model, 2026. Accessed: 2026-07-28.

Shuai Bai et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1176. URL https://aclanthology.org/2025.acl-long.1176/.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F. Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1409. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4ea4a1ea4d9ff273688c8e92bd087112-Abstract-Conference.html.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Junbo Cui et al. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026. URL https://arxiv.org/abs/2604.27393.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

Google DeepMind. Gemini 3.5 Flash Model Card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026. Published: 2026-05-19.

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. doi: 10.1109/CVPR52733.2024.01363. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024. doi: 10.1609/aaai.v38i16.29771. URL https://ojs.aaai.org/index.php/AAAI/article/view/29771.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

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 *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Meta AI. Llama 3.2 Vision Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md, 2024. Accessed: 2026-07-28.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8253–8280. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.567. URL https://aclanthology.org/2022.acl-long.567/.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. ALOHa: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 342–357. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-short.30. URL https://aclanthology.org/2024.naacl-short.30/.

Nhi Pham and Michael Schott. H-POPE: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models. *arXiv preprint arXiv:2411.04077*, 2024. URL https://arxiv.org/abs/2411.04077.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. https://qwen.ai/blog?id=qwen3.6-35b-a3b, 2026. Published: 2026-04-15.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *Computer Vision – ECCV 2024*, pp. 256–274. Springer, 2024.

doi: 10.1007/978-3-031-72943-0_15. URL https://doi.org/10.1007/978-3-031-72943-0_15.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.

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

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53553–53570. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24l.html.

Bowen Yan, Zhengsong Zhang, Liqiang Jing, Eftekhar Hossain, and Xinya Du. FIHA: Automated fine-grained hallucinations evaluations in large vision language models with davidson scene graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12014–12026. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.622. URL https://aclanthology.org/2025.findings-acl.622/.

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 30–37. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-demos.4. URL https://aclanthology.org/2022.emnlp-demos.4/.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, Huiyu Zhou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6193–6212. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.322. URL https://aclanthology.org/2025.findings-acl.322/.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao

Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

SUPPLEMENTARY MATERIAL

This document provides additional dataset details, evaluation protocols, reproducibility information, extended results, and qualitative examples supporting the main paper.

A DATASET CONSTRUCTION AND STATISTICS

A.1 SOURCE DATA AND FILTERING

CoAMA is constructed from the multi-object image pool introduced by ROPE (Chen et al., 2024), using source images from COCO (Lin et al., 2014) and ADE20K (Zhou et al., 2017). We retain images in which five visually identifiable instances can be marked and assigned instance-level class labels. Each image contains five fixed, numbered red bounding boxes, and the same object indices are used for annotation, prompting, parsing, and scoring. For each marked instance, CoAMA records five visible attributes: `color`, `material`, `shape`, `physics_state`, and `pose`. Annotators use `unknown` when the marked region does not support a reliable value; these labels remain in the annotations but are excluded from attribute scoring and Level 3 probe construction.

The five scene-composition splits are:

- `Homogeneous`: all five marked objects share one class.
- `Heterogeneous`: all five marked objects have different classes.
- `AAAAB` and `BAAAA`: four objects share one class and one differs, with the outlier appearing at different query positions.
- `Mixed`: naturally occurring class compositions not covered by the preceding templates.

Table S1: Dataset composition. Columns give images/marked instances, raw/scored/`unknown` attribute labels, and Level 3 pairs/questions.

Split	Images / instances	Attributes: raw / scored / unknown	Level 3: pairs / questions
AAAAB	102/510	2,550/2,516/34	319/638
BAAAA	100/500	2,500/2,482/18	315/630
Heterogeneous	76/380	1,900/1,868/32	366/732
Homogeneous	149/745	3,725/3,714/11	104/208
Mixed	360/1,800	9,000/8,880/120	1,431/2,862
Total	**787/3,935**	**19,675/19,460/215**	**2,535/5,070**

The 787-image benchmark contains 383 COCO images and 404 ADE20K images.

A.2 ANNOTATION PROTOCOL AND AGREEMENT

Each attribute uses a single canonical label based on the marked region. Material describes the visible exterior substance, and physical state describes the object itself, not container contents or the surrounding environment. The labels mixed and irregular are retained for composite colors and shapes. Occlusion, ambiguity, poor image quality, or insufficient evidence yields unknown. Table S2 gives the complete vocabulary. Although attribute descriptors are not guaranteed to be mutually exclusive in every context, the taxonomy is designed to provide clear visual distinctions among the properties of marked instances. Shared annotation rules make these distinctions reproducible, with exact agreement of 92.6%–98.7% across attribute types (Table S3).

Two annotators independently labeled all 3,935 marked instances using shared closed vocabularies for `color`, `material`, `shape`, `physics_state`, and `pose`. Each annotator assigned one value per attribute and used `unknown` when the marked region lacked sufficient visual evidence. Bounding boxes, object indices, and class labels were fixed before attribute annotation. The two annotation sets were matched by split, image, and object index, with identical classes and box coordinates verified for every pair. The resulting audit measures attribute-assignment reproducibility conditional on fixed object definitions; it does not evaluate localization or class-label agreement.

Table S2: Attribute taxonomy used in CoAMA.

Attribute	Allowed values
Color	red, orange, yellow, green, blue, purple, brown, white, black, gray, pink, gold, silver, transparent, mixed, unknown
Material	metal, plastic, wood, fabric, glass, ceramic, paper, stone, leather, rubber, concrete, organic, unknown
Shape	round, square, triangle, rectangular, oval, cylindrical, cubical, conical, curved, straight, flat, pointed, thin, long, irregular, unknown
Physical state	rigid, flexible, soft, liquid, gas, granular, unknown
Pose	standing, sitting, lying, hanging, leaning, tilted, upright, folded, rolled, scattered, stacked, unknown

Table S3: Attribute-wise inter-annotator agreement with `unknown` retained as a valid label. Exact agreement is the percentage of identical paired labels.

Attribute	**N**	**Exact agreement (%)**	**Cohen's κ**
Color	3,935	92.6	0.909
Material	3,935	98.0	0.974
Shape	3,935	98.0	0.967
Physical state	3,935	95.9	0.926
Pose	3,935	98.7	0.982

Tables S3 and S4 report agreement on the pre-adjudication annotations. All disagreements, including cases involving `unknown`, were subsequently resolved through adjudication before the final benchmark reference labels were fixed.

We report exact agreement and Cohen's κ separately for each attribute. For the primary analysis, `unknown` is retained as a valid categorical label, so agreement about whether an attribute is visually resolvable contributes to the reliability estimate. Labels are normalized only by lowercasing and trimming surrounding whitespace; no values are merged as synonyms. Cohen's κ is computed as $(p_o - p_e)/(1 - p_e)$, where p_o is observed agreement and p_e is the agreement expected from the annotators' empirical marginal label distributions. All statistics are computed directly from the paired human annotations; no model output enters the agreement calculation.

As a sensitivity analysis, Table S4 restricts agreement to instances for which both annotators selected a non-`unknown` value. It also reports cases in which both annotators selected `unknown` and cases in which only one annotator did so. The latter rate uses all 3,935 instances for that attribute as its denominator; these disagreements are not discarded from the primary analysis.

All five attributes achieve $\kappa > 0.90$ when `unknown` is retained. Color has the lowest agreement (92.6%, $\kappa = 0.909$), demonstrating that a pooled statistic would obscure attribute-specific variation. Material has the largest one-`unknown`/one-specific rate (1.5%); when both annotations are specific, agreement rises to 99.5% ($\kappa = 0.993$).

B Evaluation Protocol

B.1 Prompt Templates

All task instructions are provided in the user prompt. We do not add model-specific instructions beyond each model's standard input template. Each evaluation setting combines the marked image with the task question and an explicit output-format constraint.

Level 1: Multi-Object Instance Recognition

> *Given the classes: `{candidate_class_list}`. There are five red bounding boxes in the image, labeled Obj.1, Obj.2, Obj.3, Obj.4, and Obj.5. Identify the class of the object in each box. Respond exactly as: `obj1: <class1>, obj2: <class2>, obj3:`*

Table S4: Agreement when both annotators selected a specific label, followed by the distribution of `unknown` decisions. Parenthetical values are percentages; both rate columns use all 3,935 instances per attribute as the denominator.

Attribute	Both specific	Exact agreement (%)	Cohen's κ
Color	3,910	92.9	0.913
Material	3,780	99.5	0.993
Shape	3,889	98.3	0.971
Physical state	3,921	95.9	0.927
Pose	3,853	99.2	0.989

Attribute	Both unknown n (%)	One unknown, one specific n (%)
Color	10 (0.3)	15 (0.4)
Material	95 (2.4)	60 (1.5)
Shape	35 (0.9)	11 (0.3)
Physical state	11 (0.3)	3 (0.1)
Pose	60 (1.5)	22 (0.6)

<class3>, obj4: <class4>, obj5: <class5>. Do not add explanations or punctuation beyond the specified format.

Level 2: Target-Object Attribute Recognition Level 2 crosses three target-reference conditions with three response formats. The class prefix is the ground-truth class in `GT`, omitted in `None`, and replaced by the saved Level 1 prediction in `Pred`.

Single. `{class_prefix}` What is the `{attribute_type}` of Obj.`{i}`? Choose from: `{candidate_values}`. Respond with only one option.

Multi. `{class_prefix}` What are the attributes of Obj.`{i}`? Respond exactly as: `color=<value>, material=<value>, shape=<value>, physics_state=<value>, pose=<value>`.

Table. `{class_header}` For each marked object, identify its `{attribute_type}`. Respond exactly as: `Obj.1=<value>, Obj.2=<value>, Obj.3=<value>, Obj.4=<value>, Obj.5=<value>`.

The nine combinations are reported as `level2-{single|multi|table}-{gt|none|pred}`.

Level 3: Cross-Object Attribute Binding Verification All binary Level 3 settings share one attribute-aware question family. For target Obj.`{i}`, class `{c}`, attribute type `{t}`, and queried value `{v}`, the direct prompt is given below.

Obj.`{i}` is a `{c}`. `{attribute_question(i,t,v)}` Answer only yes or no.

The attribute-specific question follows three grammatical forms.

- **Color, physical state, and pose:** Is Obj.`{i} {v}`?
- **Material:** Is Obj.`{i}` made of `{v}`?
- **Shape:** Is Obj.`{i}`'s shape `{v}`?

The positive and swapped-negative probes use identical wording and differ only in `{v}`. The retained variants each modify one component of this construction.

`L3-Oracle-Direct` uses the ground-truth target class and direct object index; `L3-Oracle-Spatial` replaces the index with a deterministic spatial description (`left`, `right`, `top`, `bottom`, or `center`); and `L3-Oracle-ObjOnly` retains the index while omitting the target class.

For upstream-error diagnostics, `L3-L2GT-Direct` substitutes saved `level2-single-gt` attributes ("GT" denotes the input class for Level 2), whereas `L3-End2End-Direct` additionally substitutes the saved Level 1 class and `level2-single-pred` attributes. `L3-Multi-Attribute` queries a conjunction of two values. The remaining controls modify the evidence. `L3-PresentVsAbsent` contrasts source-present and marked-set-absent negatives, `L3-TargetCrop` crops around the target, and `L3-SourceMasked` masks the source.

For `L3-Multi-Attribute`, the direct question becomes the following.

> *Obj.`{i}` is a `{c}`. Is Obj.`{i}` `{attribute_value_1}` and `{attribute_value_2}`? Answer only yes or no.*

B.2 Level 3 Pair Construction

Oracle settings use ground-truth target classes and annotated attribute values to specify the queries, reducing the influence of upstream prediction errors on query construction. This design establishes instance-level attribute binding as a distinct evaluation target. The selected probes are evaluated regardless of the model's corresponding Level 1 or Level 2 predictions. The three levels use different response spaces, scoring rules, and evaluation sets and should not be interpreted as a monotonic ordering of task difficulty. The released `level3_oracle_manifest.json` specifies the 2,535 positive–swap pairs (5,070 questions) used by L3-Oracle-Direct. A candidate has distinct target and source instances with unequal, scorable canonical values of the same attribute type. The saved manifest contains at most one selected pair per image and attribute type; the overall score weights selected pairs equally, not all possible ordered target–source combinations. Different source instances therefore do not contribute repeated positives for the same image–attribute unit in this manifest.

Reproduction of the reported Direct evaluation uses these saved probes rather than resampling candidates. The released sampler enumerates eligible ordered candidates and uses a random choice for each image–attribute unit, whereas the frozen file fixes the realized choices. Its content hash and per-split counts provide the reproducible specification of the evaluated set.

The visual-evidence controls modify the image while preserving the original target–source definition. The error-propagation variants inject predictions from earlier levels to quantify how upstream errors affect binding verification.

B.3 L3-PresentVsAbsent Construction

A marked-set-absent candidate belongs to the queried attribute taxonomy, is not assigned to any of the five marked objects, and is neither unknown, mixed, nor irregular. The source-present branch instead uses a value annotated on another marked instance. For example, a brown dining table may be queried with green from a marked broccoli and with red as a marked-set-absent alternative; an orange carrot may similarly be queried with green and gray.

The released sampler chooses an absent value by an index computed from the image identifier, attribute type, and target index. The saved questions specify the values actually evaluated. The paired comparison estimates the difference between the two saved negative branches. Absence is defined over marked-object annotations and may not hold in unmarked regions.

B.4 Metrics

For a complete positive–swap pair, PairAcc requires yes on the positive branch and no on the swapped branch. It is micro-averaged over complete scored pairs. $\mathrm{TAR}_{\mathrm{pos}}$, $\mathrm{FAR}_{\mathrm{swap}}$, and $\mathrm{FAR}_{\mathrm{absent}}$ are the proportions of yes responses among all questions in the positive, swapped-negative, and absent-negative branches, respectively. In symbols, each branch rate is $N_{\mathrm{yes}}/N_{\mathrm{branch}}$, using that branch's own evaluated set.

YesRatio divides the number of yes responses by all evaluated questions in the reported setting; parse-failure rate uses the same total question count. Invalid answers remain in all applicable denominators and count as incorrect for accuracy and PairAcc, but not as yes for acceptance rates. These conventions prevent a low FAR caused by unparsable outputs from being interpreted without the accompanying parse-failure rate.

Table S5: Evaluated models and decoding configurations.

Display name	Evaluated identifier	Access	Precision / execution	Decoding
Llama-3.2-11B	`meta-llama/Llama-3.2-11B-Vision-Instruct`	Local checkpoint	BF16, batch 1	Greedy, max 500 new tokens
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`	Local checkpoint	Automatic supported dtype, batch 1	Greedy, max 500 new tokens
Qwen3-VL-32B	`Qwen/Qwen3-VL-32B-Instruct`	Local checkpoint	Automatic supported dtype, batch 1	Greedy, max 500 new tokens
Qwen3.6-35B-A3B	`Qwen/Qwen3.6-35B-A3B`	Local checkpoint	BF16-compatible execution, batch 1	Greedy, max 500 new tokens
Qwen3-VL-32B-Thinking	`Qwen/Qwen3-VL-32B-Thinking`	Local checkpoint	Automatic supported dtype, batch 1; thinking enabled	Greedy, max 4,096 new tokens; final-answer parsing
MiniCPM-o 4.5	`openbmb/MiniCPM-o-4_5`	Local checkpoint	BF16, batch 1; vision mode; thinking disabled	Greedy, max 500 new tokens
Gemini 3.5 Flash	`Gemini-3.5-Flash`	API access, 21–24 July 2026	Provider-managed	Temperature 0, max 500 output tokens
Qwen3.7-Plus	`Qwen3.7-Plus`	API access, 24 July 2026	Provider-managed	Temperature 0, max 500 output tokens

B.5 Parsing and Scoring

Level 1 parsing extracts `Obj.1` through `Obj.5` case-insensitively, removes formatting wrappers and terminal punctuation, and applies exact match to normalized lowercase class strings. Level 2 normalization lowercases outputs, removes Markdown emphasis and surrounding quotation or punctuation marks, converts internal whitespace to underscores, and accepts only values in the predefined taxonomy. Multi-attribute and table responses must expose the expected key-value structure. Ground-truth `unknown` labels are excluded before scoring.

Level 3 parsing searches for word-boundary `yes` and `no` tokens in the model-specific final-answer region. A response is valid only when exactly one polarity is recoverable. Responses containing neither polarity or both polarities are parse failures. Format-aware extraction handles models that emit a dedicated `FINAL` or answer field, but the semantic decision is never manually corrected.

Parse failures remain in the denominator and are scored as incorrect. Responses are not manually edited or repaired, and semantic failures are not requeried. Infrastructure retries are limited to connection errors, timeouts, HTTP 429 responses, and server-side 5xx errors; the final successful response is retained.

C Experimental Setup and Statistical Procedure

C.1 Model Configurations

Images are passed through each model's native processor. The benchmark does not impose a global resize beyond model-native preprocessing. All main runs use the original marked PNG; only the explicitly named visual-intervention settings crop or mask pixels.

C.2 Compute Environment and Execution Policy

Open-weight models were evaluated on two NVIDIA A800 GPUs, each with 80 GB of memory, using two local container configurations. Llama-3.2-11B, MiniCPM-o 4.5, and the Qwen3-VL mod-

els were evaluated in the standard container, `moh-new-vlm`. The reference software environment comprised Ubuntu 22.04, Python 3.11, PyTorch 2.6.0, CUDA 12.4.1, Transformers 4.57.0, Accelerate 1.13.0, and `qwen-vl-utils` 0.0.14. Qwen3.6-35B-A3B was evaluated in the dedicated GCC-enabled container, `moh-new-vlm`, with model-specific parameters configured in accordance with its official documentation. All models were evaluated with a batch size of 1 and their native image processors. Greedy decoding with a temperature of 0 was applied whenever the model interface exposed decoding controls. Each query was evaluated once. Transport-level failures were retried up to three times, whereas parsing failures were retained as-is without reissuing the corresponding queries.

C.3 Statistical Uncertainty

We compute 95% confidence intervals using 2,000 percentile bootstrap replicates clustered by image. Paired comparisons preserve model and pair identity before image-level resampling. Statistical resampling uses seed `20260724`.

D Additional Level 1 and Level 2 Results

D.1 Level 1 Recognition

Table S6: Level 1 multi-object recognition. Acc. and PF are percentages; PF denotes parse-failure rate, and `N` is the number of marked instances. Confidence intervals use image-cluster bootstrap.

Model	Acc. (%)	95% CI
Llama-3.2-11B	72.53	[70.47, 74.51]
Qwen3-VL-8B	72.66	[70.67, 74.64]
Qwen3-VL-32B	75.30	[73.39, 77.18]
Qwen3.6-35B-A3B	77.08	[75.20, 78.88]
Qwen3-VL-32B Thinking	69.78	[67.52, 72.02]
MiniCPM-o 4.5	61.47	[59.26, 63.66]
Gemini 3.5 Flash	85.64	[84.14, 87.12]
Qwen3.7-Plus	81.37	[79.64, 83.08]

Model	PF (%)	N
Llama-3.2-11B	4.85	3,935
Qwen3-VL-8B	0.38	3,935
Qwen3-VL-32B	0.08	3,935
Qwen3.6-35B-A3B	0.00	3,935
Qwen3-VL-32B Thinking	7.90	3,935
MiniCPM-o 4.5	0.03	3,935
Gemini 3.5 Flash	0.00	3,935
Qwen3.7-Plus	0.00	3,935

D.2 Complete Level 2 Matrix

Table S7: Complete Level 2 attribute-recognition matrix, organized by output format. Models occupy rows; each label condition reports accuracy (Acc.), image-cluster bootstrap 95% confidence interval, parse-failure rate (PF), and Count. Count is the number of scored attribute responses used as the common denominator for Acc. and PF; Acc., confidence bounds, and PF are percentages. API-accessed models were evaluated only under the Single-GT setting; dashes denote settings that were not run.

Model	Acc.	95% CI	PF	Count
Single-GT (%)				
Llama-3.2-11B	47.49	[46.20,48.80]	0.18	19,460
Qwen3-VL-8B	53.05	[51.69,54.43]	0.12	19,460
Qwen3-VL-32B	56.06	[54.82,57.32]	0.54	19,460
Qwen3.6-35B-A3B	55.22	[53.91,56.55]	0.68	19,460
Qwen3-VL-32B Thinking	62.84	[61.77,63.92]	1.21	19,460
MiniCPM-o 4.5	50.05	[48.63,51.52]	2.09	19,460
Gemini 3.5 Flash	66.56	[65.60,67.52]	4.58	19,460
Qwen3.7-Plus	64.58	[63.66,65.52]	2.30	19,460
Single-None (%)				
Llama-3.2-11B	42.13	[40.80,43.50]	0.30	19,460
Qwen3-VL-8B	43.37	[41.93,44.89]	0.15	19,460
Qwen3-VL-32B	47.88	[46.49,49.24]	0.57	19,460
Qwen3.6-35B-A3B	48.48	[47.17,49.83]	2.07	19,460
Qwen3-VL-32B Thinking	51.76	[50.51,53.04]	1.88	19,460
MiniCPM-o 4.5	38.42	[36.99,39.84]	2.20	19,460
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Single-Pred (%)				
Llama-3.2-11B	44.05	[42.63,45.40]	4.94	19,460
Qwen3-VL-8B	48.65	[47.19,50.07]	0.48	19,460
Qwen3-VL-32B	52.28	[50.94,53.53]	0.38	19,460
Qwen3.6-35B-A3B	51.87	[50.53,53.21]	0.46	19,460
Qwen3-VL-32B Thinking	53.85	[52.22,55.40]	8.89	19,460
MiniCPM-o 4.5	44.41	[43.00,45.81]	0.69	19,460
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Multi-GT (%)				
Llama-3.2-11B	44.06	[42.85,45.25]	5.88	19,460
Qwen3-VL-8B	57.10	[55.85,58.36]	1.58	19,460
Qwen3-VL-32B	61.96	[60.82,63.16]	1.59	19,460
Qwen3.6-35B-A3B	53.59	[52.28,54.88]	5.55	19,460
Qwen3-VL-32B Thinking	66.74	[65.64,67.77]	0.96	19,460
MiniCPM-o 4.5	49.95	[48.55,51.43]	14.94	19,460
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Multi-None (%)				
Llama-3.2-11B	42.24	[40.20,44.34]	6.30	5,933
Qwen3-VL-8B	48.15	[46.00,50.17]	1.79	5,933
Qwen3-VL-32B	53.55	[51.43,55.79]	1.47	5,933
Qwen3.6-35B-A3B	47.02	[45.68,48.29]	7.45	19,460
Qwen3-VL-32B Thinking	59.36	[57.48,61.40]	1.01	5,933
MiniCPM-o 4.5	37.30	[34.85,39.66]	21.84	5,933
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Multi-Pred (%)				
Llama-3.2-11B	43.11	[41.13,45.26]	9.67	5,933

Table S7 (continued)

Model	Acc.	95% CI	PF	Count
Qwen3-VL-8B	52.96	[50.59,55.29]	2.56	5,933
Qwen3-VL-32B	59.40	[57.16,61.54]	1.67	5,933
Qwen3.6-35B-A3B	50.63	[49.30,51.92]	5.38	19,460
Qwen3-VL-32B Thinking	58.79	[55.95,61.63]	7.58	5,933
MiniCPM-o 4.5	46.47	[43.97,48.82]	11.85	5,933
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Table-GT (%)				
Llama-3.2-11B	26.56	[25.08,27.98]	45.67	19,460
Qwen3-VL-8B	55.36	[54.00,56.71]	0.53	19,460
Qwen3-VL-32B	58.65	[57.32,59.99]	0.73	19,460
Qwen3.6-35B-A3B	58.98	[57.72,60.17]	0.62	19,460
Qwen3-VL-32B Thinking	64.46	[63.46,65.48]	1.19	19,460
MiniCPM-o 4.5	53.35	[52.00,54.74]	5.19	19,460
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Table-None (%)				
Llama-3.2-11B	27.71	[25.41,30.24]	38.70	5,933
Qwen3-VL-8B	46.97	[44.49,49.39]	0.44	5,933
Qwen3-VL-32B	49.81	[47.33,52.38]	0.27	5,933
Qwen3.6-35B-A3B	49.95	[48.61,51.29]	0.37	19,460
Qwen3-VL-32B Thinking	53.40	[51.13,55.68]	2.81	5,933
MiniCPM-o 4.5	40.08	[37.75,42.44]	5.21	5,933
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Table-Pred (%)				
Llama-3.2-11B	21.17	[18.67,23.88]	55.45	5,933
Qwen3-VL-8B	51.27	[48.96,53.73]	1.30	5,933
Qwen3-VL-32B	54.44	[51.94,56.87]	1.11	5,933
Qwen3.6-35B-A3B	55.14	[53.86,56.43]	0.56	19,460
Qwen3-VL-32B Thinking	56.38	[53.46,59.20]	8.12	5,933
MiniCPM-o 4.5	46.69	[44.32,49.08]	3.07	5,933
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–

The label conditions reveal a consistent role for class semantics. Under the Single format, `GT` exceeds `None` for all six open-weight models by 5.36–11.63 percentage points. Supplying the target class narrows the plausible attribute space and links the object index to a semantic category, whereas `None` requires the model to resolve the target solely from the visual marker and local evidence. `Pred` lies between `GT` and `None` for every model in the Single setting. Predicted classes therefore preserve useful semantic information, but recognition errors weaken the prior and propagate to attribute prediction.

The three output formats do not form a universal difficulty ordering. Relative to Single-GT, Multi-GT improves Qwen3-VL-8B, Qwen3-VL-32B, and Qwen3-VL-32B-Thinking by 4.05, 5.90, and 3.90 points, respectively. Table-GT also improves Qwen3.6-35B-A3B and MiniCPM-o 4.5 by 3.76 and 3.30 points. Joint queries can provide a regular response structure that some models exploit effectively, but they also require the model to preserve field names, object indices, and attribute ownership across a longer output.

Llama-3.2-11B is the clearest format-sensitive case. Its accuracy decreases from 47.49% on Single-GT to 44.06% on Multi-GT and 26.56% on Table-GT. At the same time, parse failure rises from 0.18% to 5.88% and 45.67%, respectively; Table-Pred reaches a 55.45% parse-failure rate. The table condition therefore measures both attribute recognition and the ability to maintain a complete

Table S8: Level 2 `Single-GT` by attribute. Each cell reports `Acc. / PF`, both in percent; `N` is shared across models.

Model	Color (N=3,923)	Material (N=3,839)	Shape (N=3,899)	Physical state (N=3,924)	Pose (N=3,875)
Llama-3.2-11B	44.74 / 0.00	44.26 / 0.31	33.83 / 0.56	58.82 / 0.00	55.74 / 0.03
Qwen3-VL-8B	47.59 / 0.15	61.58 / 0.23	45.88 / 0.05	61.37 / 0.03	48.90 / 0.15
Qwen3-VL-32B	53.48 / 0.28	67.54 / 0.91	43.73 / 0.28	62.54 / 0.25	53.16 / 0.98
Qwen3.6-35B-A3B	55.65 / 0.20	66.11 / 1.51	41.75 / 0.15	62.28 / 0.00	50.40 / 1.57
Qwen3-VL-32B-Thinking	53.68 / 1.10	81.64 / 3.41	57.68 / 0.21	71.30 / 0.97	50.12 / 0.39
MiniCPM-o 4.5	42.39 / 1.48	59.55 / 4.19	40.04 / 0.33	60.91 / 0.82	47.46 / 3.66
Gemini 3.5 Flash	53.45 / 4.74	86.25 / 1.80	57.55 / 10.18	74.64 / 1.02	61.19 / 5.14
Qwen3.7-Plus	53.66 / 1.15	84.42 / 1.56	59.14 / 4.74	65.77 / 1.48	60.26 / 2.58

Table S9: Level 2 `Single-GT` by scene-composition split. Each cell reports `Acc. / PF`, both in percent; `N` is the number of scorable labels in the column.

Model	Homogeneous (N=3,714)	Heterogeneous (N=1,868)	AAAAB (N=2,516)	BAAAA (N=2,482)	Mixed (N=8,880)
Llama-3.2-11B	44.02 / 0.32	51.93 / 0.16	52.74 / 0.32	47.34 / 0.20	46.55 / 0.08
Qwen3-VL-8B	48.38 / 0.11	59.64 / 0.00	56.56 / 0.24	53.26 / 0.04	52.56 / 0.15
Qwen3-VL-32B	53.26 / 0.24	60.87 / 1.07	58.66 / 0.36	55.08 / 0.52	55.77 / 0.61
Qwen3.6-35B-A3B	51.72 / 0.35	60.92 / 1.39	59.06 / 0.40	54.83 / 0.36	54.50 / 0.84
Qwen3-VL-32B-Thinking	63.03 / 0.62	63.12 / 1.77	66.18 / 1.23	65.55 / 1.09	61.00 / 1.36
MiniCPM-o 4.5	46.88 / 0.57	53.05 / 4.07	52.34 / 1.39	49.68 / 2.34	50.19 / 2.43
Gemini 3.5 Flash	67.99 / 4.93	67.02 / 2.89	68.92 / 4.77	66.84 / 5.56	65.11 / 4.46
Qwen3.7-Plus	62.33 / 2.69	68.52 / 1.93	65.86 / 1.99	63.66 / 1.85	64.59 / 2.43

multi-object schema. For this model, much of the apparent Table degradation is attributable to format compliance rather than a proportional loss of visual attribute knowledge.

Together, the matrix shows that Level 2 scores reflect three interacting capabilities, namely visual attribute recovery, class-semantic support, and output organization. Accuracy and parse failure are consequently reported together, and format-specific scores should not be interpreted as a pure ordering of visual difficulty.

D.3 Single-GT Breakdowns

Attribute-wise results show that the aggregate Level 2 score is not driven uniformly by the five taxonomies. Material is the strongest Single-GT attribute for six of the eight models, reaching 81.64% for Qwen3-VL-32B-Thinking, 86.25% for Gemini 3.5 Flash, and 84.42% for Qwen3.7-Plus. Physical state is strongest for Llama and MiniCPM. Shape is the weakest attribute for four open-weight models, while color is the weakest for both API-accessed models. These differences are consistent with the distinct evidence required by each attribute. Material and physical state draw on visible surface and structural properties, whereas color and shape demand more localized appearance evidence.

Parse behavior also varies by attribute. Gemini attains the highest material accuracy but records 10.18% parse failure on shape, demonstrating that strong aggregate recognition can coexist with attribute-specific output instability. The per-attribute table therefore separates semantic difficulty from invalid-response effects that would be hidden by a single overall score.

Table S10: L3-Oracle-Direct on 2,535 shared pairs. `PairAcc` requires correct positive and swapped-negative decisions. FAR and TAR denote `FAR_swap` and `TAR_pos`; `YesRatio` is the proportion of all questions whose parsed answer is yes.

Model	PairAcc (%)	95% CI	FAR (%)	TAR (%)	YesRatio (%)	PF (%)	Pairs
Qwen3.7-Plus	67.69	[65.88, 69.51]	15.82	78.93	47.38	0.22	2,535
Qwen3.6-35B-A3B	63.55	[61.71, 65.44]	17.28	75.19	46.23	0.00	2,535
Gemini 3.5 Flash	61.58	[59.70, 63.46]	9.66	68.44	39.05	0.99	2,535
Qwen3-VL-32B-Thinking	61.03	[59.18, 62.81]	18.26	72.58	45.42	0.00	2,535
Qwen3-VL-32B	58.19	[56.22, 60.16]	20.99	73.25	47.12	0.00	2,535
Qwen3-VL-8B	57.20	[55.26, 59.04]	21.14	71.12	46.13	0.00	2,535
MiniCPM-o 4.5	55.58	[53.63, 57.55]	23.51	72.86	48.19	0.00	2,535
Llama-3.2-11B	43.04	[41.03, 45.13]	47.53	84.22	65.88	0.02	2,535

Table S11: L3-Oracle-Direct PairAcc by attribute type. All entries are percentages.

Model	Color	Material	Shape	Physical state	Pose
Llama-3.2-11B	49.05	64.02	34.95	29.66	31.09
Qwen3-VL-8B	53.48	69.19	55.09	67.16	43.19
Qwen3-VL-32B	55.22	68.82	55.79	68.14	44.91
Qwen3.6-35B-A3B	57.75	74.54	61.81	76.23	50.67
Qwen3-VL-32B-Thinking	56.33	70.30	63.43	73.53	45.30
MiniCPM-o 4.5	43.67	72.32	56.25	72.30	38.96
Gemini 3.5 Flash	54.27	71.22	60.88	75.49	50.10
Qwen3.7-Plus	58.86	74.35	71.06	87.25	53.36
Cross-model mean	**53.58**	**70.60**	**57.41**	**68.72**	**44.70**

Scene composition produces a related but non-universal pattern. Homogeneous is the lowest-accuracy split for six of eight models, including five of the six open-weight models. Repeating the same class across all marked objects reduces the discriminative value of the class prefix and increases reliance on the numbered visual markers. Qwen3-VL-32B-Thinking and Gemini instead reach their lowest point estimates on Mixed, showing that class repetition is not the only source of Level 2 difficulty. Across all splits, the smaller variation at Level 2 than the same-category gaps observed at Level 3 indicates that recognizing an attribute and assigning it to the correct instance remain distinguishable capabilities.

E Additional Level 3 Results

E.1 L3-Oracle-Direct

Cross-object binding remains challenging across the evaluated models, as the highest `PairAcc` is 67.69%. The component metrics clarify why a paired score is necessary. Llama attains the highest `TAR_pos` (84.22%) but also the highest `FAR_swap` (47.53%), yielding the lowest `PairAcc`. Gemini shows the lowest `FAR_swap` (9.66%) but also a lower `TAR_pos` (68.44%) than most models. Reliable binding therefore requires both recognizing the target attribute and rejecting a source-derived value grounded elsewhere in the image.

E.2 Attribute Breakdown

Pose is the most difficult attribute type, with a cross-model mean `PairAcc` of 44.70%. Color and shape form the next weakest group at 53.58% and 57.41%. These attributes depend strongly on localized appearance, contour, orientation, and viewpoint, making it necessary to associate fine-grained evidence with one marked instance rather than the scene as a whole.

Table S12: L3-Oracle-Direct PairAcc by scene-composition split. All entries are percentages.

Model	Homogeneous	Heterogeneous	AAAAB	BAAAA	Mixed
Llama-3.2-11B	40.38	43.99	41.38	42.86	43.40
Qwen3-VL-8B	44.23	56.83	52.98	54.92	59.68
Qwen3-VL-32B	46.15	57.38	56.74	56.51	59.96
Qwen3.6-35B-A3B	48.08	63.39	63.64	64.76	64.43
Qwen3-VL-32B-Thinking	49.04	62.57	58.93	58.10	62.61
MiniCPM-o 4.5	34.62	57.65	56.11	50.79	57.51
Gemini 3.5 Flash	50.96	62.02	65.20	60.95	61.57
Qwen3.7-Plus	54.81	67.21	70.53	68.57	67.92

Table S13: Same-category versus different-category target-source pairs. There are 217 same-category and 2,318 different-category pairs.

Model	Same-category PairAcc (%)	Different-category PairAcc (%)	Same - different (pp)	95% CI (pp)
Llama-3.2-11B	33.64	43.92	-10.28	[-16.94, -3.35]
Qwen3-VL-8B	42.86	58.54	-15.68	[-22.35, -8.80]
Qwen3-VL-32B	41.01	59.79	-18.78	[-25.54, -12.18]
Qwen3.6-35B-A3B	49.77	64.84	-15.07	[-22.42, -7.84]
Qwen3-VL-32B-Thinking	46.08	62.42	-16.34	[-23.55, -9.36]
MiniCPM-o 4.5	30.41	57.94	-27.52	[-33.88, -20.69]
Gemini 3.5 Flash	52.53	62.42	-9.89	[-17.04, -2.59]
Qwen3.7-Plus	54.38	68.94	-14.56	[-21.63, -7.40]

Material and physical state are more reliably bound on average, reaching 70.60% and 68.72%, but the model profiles remain distinct. Llama reaches only 29.66% on physical state because it accepts 68.87% of swapped physical-state negatives, whereas Qwen3.7-Plus reaches 87.25%. The attribute taxonomy therefore exposes different binding behaviors rather than a single global difficulty scale.

E.3 Scene Composition and Instance Relations

Homogeneous is the most difficult split, with a cross-model mean `PairAcc` of 46.03%; seven of eight models obtain their lowest split-level result in this condition. When all five marked objects share a class, class semantics cannot distinguish the target from its competitors, so the model must rely more heavily on the visual marker and instance-local evidence. This pattern is consistent with the direct same-category comparison below.

The AAAAB and BAAAA results do not exhibit a shared ordering. Moving the outlier class to a different query position affects models differently, suggesting an interaction between scene composition and model-specific response organization. Mixed scenes have the highest descriptive mean, while Qwen3.7-Plus reaches its strongest split result on AAAAB.

All eight models perform worse on same-category pairs. The cross-model mean gap is -16.02 percentage points, and every confidence interval lies below zero. When the target and source share the same class, their class labels alone do not identify which instance owns the queried attribute. The result highlights the need to preserve attribute ownership between individual marked instances.

MiniCPM-o 4.5 shows the largest gap, decreasing from 57.94% to 30.41%. Qwen3.7-Plus attains the highest same-category `PairAcc`, yet its 54.38% result remains 14.56 points below its different-category performance.

Target size is measured as normalized bounding-box area. Common tertile thresholds are 0.01877 and 0.04214; the contrast uses 845 large and 845 small targets and omits the middle tertile. The

Table S14: Target-size and source-distance diagnostics.

Model	Large - Small PairAcc (pp)	95% CI (pp)	Far - Near PairAcc (pp)	95% CI (pp)
Llama-3.2-11B	-2.84	[-7.82, +2.17]	-0.45	[-4.15, +3.24]
Qwen3-VL-8B	-1.66	[-6.56, +3.27]	-1.60	[-5.67, +2.40]
Qwen3-VL-32B	-2.49	[-7.17, +2.14]	+0.85	[-2.99, +4.83]
Qwen3.6-35B-A3B	-3.31	[-7.91, +1.19]	-3.24	[-7.12, +0.32]
Qwen3-VL-32B-Thinking	-5.33	[-10.15, -0.44]	-1.51	[-5.04, +2.03]
MiniCPM-o 4.5	-0.83	[-5.74, +3.94]	+1.31	[-2.51, +5.11]
Gemini 3.5 Flash	-8.52	[-13.11, -4.09]	-1.04	[-4.67, +2.46]
Qwen3.7-Plus	-3.55	[-8.10, +0.97]	-0.79	[-4.42, +2.81]

Table S15: Absolute metrics for L3-TargetCrop and L3-SourceMasked. FAR and TAR denote `FAR_swap` and `TAR_pos`; `YesRatio` follows Table S10, and PF denotes parse-failure rate. Each row contains 2,535 pairs and 5,070 questions.

Model	Setting	PairAcc (%) ↑	FAR (%) ↓	TAR (%) ↑	YesRatio (%)	PF (%) ↓
Llama-3.2-11B	L3-TargetCrop	42.96	46.55	84.26	65.40	0.63
Llama-3.2-11B	L3-SourceMasked	43.43	44.50	82.09	63.29	0.14
Qwen3-VL-8B	L3-TargetCrop	61.34	21.70	77.16	49.43	0.00
Qwen3-VL-8B	L3-SourceMasked	57.20	20.99	71.16	46.07	0.00
Qwen3-VL-32B	L3-TargetCrop	60.16	20.67	75.27	47.97	0.00
Qwen3-VL-32B	L3-SourceMasked	58.82	21.22	74.44	47.83	0.00
Qwen3.6-35B-A3B	L3-TargetCrop	64.89	18.11	77.79	47.95	0.00
Qwen3.6-35B-A3B	L3-SourceMasked	63.19	17.55	75.19	46.37	0.00
Qwen3-VL-32B-Thinking	L3-TargetCrop	61.11	16.80	72.43	44.62	0.00
Qwen3-VL-32B-Thinking	L3-SourceMasked	61.78	17.99	73.29	45.64	0.02
MiniCPM-o 4.5	L3-TargetCrop	61.26	22.68	78.58	50.63	0.00
MiniCPM-o 4.5	L3-SourceMasked	55.78	23.31	71.91	47.61	0.00

Large-minus-Small estimate is negative for every model, but only Qwen3-VL-32B-Thinking and Gemini have intervals entirely below zero. Larger targets therefore do not provide a general binding advantage.

Source distance is normalized center distance, split at the common median 0.31948 into 1,266 far and 1,269 near pairs. Far-minus-Near differences are small, change direction across models, and have intervals that all include zero. Spatial proximity is thus a weaker diagnostic factor than source-target category identity and scene composition.

E.4 Other Diagnostics

Table S15 reports the absolute metrics for target cropping and source masking. TargetCrop improves PairAcc for five of six models, raising the mean from 56.43% to 58.62%, whereas SourceMasked

yields a mean of 56.70%. Thus, focusing visual input on the target provides a larger observed benefit than removing the source alone. These controls reveal the role of visual-evidence availability in binding performance, consistent with Section 4.3. The remaining analysis combines target-reference variants, upstream error propagation, and multi-attribute queries.

Table S16 separates three factors that are aggregated in end-to-end binding: how the target is referenced, whether upstream predictions replace oracle information, and whether two attributes are queried jointly. Oracle-Direct covers all eight models; the remaining diagnostic settings cover the six open-weight models for which the complete matrix is available.

Table S16: Level 3 diagnostic matrix. Models occupy rows; each setting reports PairAcc, false-acceptance rate (FAR), true-acceptance rate (TAR), and parse-failure rate (PF), all in percent. Oracle-Direct, Oracle-Spatial, and Oracle-ObjOnly use 2,535 pairs; Multi-Attribute uses 2,452 pairs. L2GT-Direct and End2End-Direct use each model's eligible upstream-prediction subset. Dashes denote settings not run for the API-accessed models.

Model	PairAcc	FAR	TAR	PF
Oracle-Direct (%)				
Llama-3.2-11B	43.04	47.53	84.22	0.02
Qwen3-VL-8B	57.20	21.14	71.12	0.00
Qwen3-VL-32B	58.19	20.99	73.25	0.00
Qwen3.6-35B-A3B	63.55	17.28	75.19	0.00
Qwen3-VL-32B Thinking	61.03	18.26	72.58	0.00
MiniCPM-o 4.5	55.58	23.51	72.86	0.00
Gemini 3.5 Flash	61.58	9.66	68.44	0.99
Qwen3.7-Plus	67.69	15.82	78.93	0.22
Oracle-Spatial (%)				
Llama-3.2-11B	38.30	37.51	69.86	4.64
Qwen3-VL-8B	51.68	29.98	73.29	0.00
Qwen3-VL-32B	52.66	28.56	74.32	0.00
Qwen3.6-35B-A3B	57.36	23.71	74.08	0.00
Qwen3-VL-32B Thinking	56.69	21.74	71.08	0.00
MiniCPM-o 4.5	48.40	31.01	71.52	0.00
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Oracle-ObjOnly (%)				
Llama-3.2-11B	37.75	45.56	69.86	0.02
Qwen3-VL-8B	39.49	38.26	59.92	0.00
Qwen3-VL-32B	45.36	34.95	64.42	0.00
Qwen3.6-35B-A3B	50.81	22.33	60.04	0.00
Qwen3-VL-32B Thinking	48.95	29.19	64.38	0.02
MiniCPM-o 4.5	35.98	46.23	62.64	0.00
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
L2GT-Direct (%)				
Llama-3.2-11B	34.45	57.92	82.76	0.04
Qwen3-VL-8B	42.08	36.77	64.34	0.00
Qwen3-VL-32B	46.52	33.92	70.54	0.00
Qwen3.6-35B-A3B	49.55	32.67	69.98	0.00
Qwen3-VL-32B Thinking	45.03	33.96	65.97	0.00
MiniCPM-o 4.5	46.01	35.45	69.64	0.00
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
End2End-Direct (%)				
Llama-3.2-11B	31.53	61.41	80.28	0.02
Qwen3-VL-8B	36.50	41.95	58.70	0.00

Table S16 (continued)

Model	PairAcc	FAR	TAR	PF
Qwen3-VL-32B	44.08	40.87	69.57	0.00
Qwen3.6-35B-A3B	43.81	36.83	64.55	0.00
Qwen3-VL-32B Thinking	40.53	38.98	60.41	0.02
MiniCPM-o 4.5	35.60	45.50	59.17	0.00
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–
Multi-Attribute (%)				
Llama-3.2-11B	49.76	45.96	93.88	0.65
Qwen3-VL-8B	61.01	13.25	68.92	0.00
Qwen3-VL-32B	68.92	11.79	77.77	0.00
Qwen3.6-35B-A3B	70.39	7.30	75.33	0.00
Qwen3-VL-32B Thinking	67.90	8.93	74.14	0.00
MiniCPM-o 4.5	61.79	9.91	68.11	0.00
Gemini 3.5 Flash	–	–	–	–
Qwen3.7-Plus	–	–	–	–

Replacing the direct object-class reference with a spatial description lowers `PairAcc` for every model. Removing the class name altogether produces a further decline, with L3-Oracle-ObjOnly ranging from 35.98% to 50.81%. The target class therefore provides useful semantic support for resolving which marked instance is being queried.

Upstream prediction errors compound the binding problem. For every model, performance decreases from L3-Oracle-Direct to L3-L2GT-Direct and again to L3-End2End-Direct. Injecting predicted Level 2 attributes reduces `PairAcc` to 34.45%–49.55%; additionally replacing the target class with the Level 1 prediction lowers it to 31.53%–44.08%. Together, the monotonic declines across all six models demonstrate the propagation of upstream prediction errors into binding judgments and establish upstream recognition quality as a key factor in reliable attribute ownership verification.

L3-Multi-Attribute has a higher `PairAcc` than L3-Oracle-Direct for all six models, primarily because its conjunction queries reduce false acceptance. It is therefore a distinct compositional diagnostic rather than a strictly harder version of the single-attribute swap.

F Qualitative Examples

Table S17 presents three cases illustrating distinct model behaviors on scenes without marked human subjects. Object indices refer to the red boxes in each image.

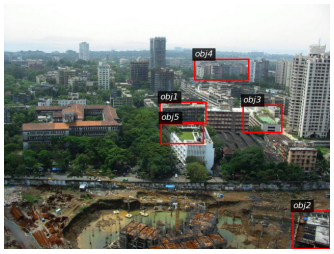
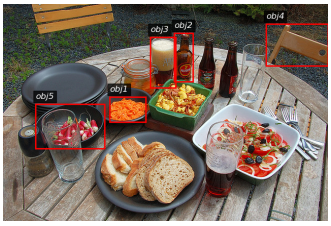
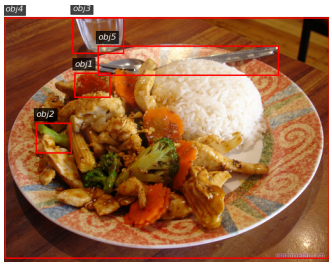
F.1 Attribute Transfer Between Same-Category Instances

Both target and source share the class `building`. The model accepts the target value but fails to reject the color transferred from another marked building. This example illustrates the same-category gap in Table S13, where the shared class label does not distinguish the two instances' attribute ownership.

F.2 Cropping Can Restore Missed Target Evidence

The direct prompt already rejects the swapped value but misses the bottle's true material. Cropping preserves that rejection while recovering the positive judgment. This example shows how target-focused evidence can improve `PairAcc` through `TAR_pos`, consistent with the component-level interpretation of Table S15.

Table S17: Qualitative examples. The left case transfers a source color between same-class instances; the center case recovers target evidence through cropping; the right case accepts a negative absent from the marked-object annotations. Response pairs list the positive and negative judgments in that order.

Same-category wrong binding	Target-crop recovery	Absent-negative false acceptance
		
Setting: Homogeneous; L3-Oracle-Direct **Model:** Llama-3.2-11B **Target / source:** Obj.1 building / Obj.3 building **Target / swapped value:** `brown` / `mixed` **Responses:** Yes / Yes **Pair outcome:** Incorrect	**Setting:** Oracle Direct vs. TargetCrop **Model:** Llama-3.2-11B **Target / source:** Obj.2 bottle / Obj.5 bowl **Target / swapped value:** `glass` / `ceramic` **Oracle Direct:** No / No (Incorrect) **TargetCrop:** Yes / No (Correct)	**Setting:** L3-PresentVsAbsent **Model:** Llama-3.2-11B **Target / matched source:** Obj.4 dining table / Obj.2 broccoli **Target color:** `brown` **Source-present / marked-set-absent:** `green` / `red` **Absent-negative response:** Yes (False acceptance)

F.3 Marked-Set Absence Does Not Guarantee Rejection

Although `red` is not assigned to any marked object, the model still accepts it for the dining table. This case shows that marked-set absence alone does not guarantee rejection. As throughout the benchmark, marked-set absence does not assert pixel-level absence from every unannotated region.

G Scope and Limitations

CoAMA evaluates static multi-object scenes through a defined taxonomy of explicit visual attributes and controlled binary questions. Performance may vary with prompt wording, decoding strategy, language, and visual domain. Future extensions can broaden the framework to relations, temporal changes, video, embodied interaction, multilingual evaluation, and open-world attributes.

Although natural-language attribute descriptors are not universally mutually exclusive, CoAMA emphasizes clear visual distinctions under shared instance-level annotation rules. Its canonical labels specify the property assigned to each marked object, and the high agreement in Section A.2 supports consistent identification of these distinctions. Oracle queries use ground-truth annotations to reduce the influence of upstream prediction errors on query construction, supporting targeted evaluation of instance-level attribute binding (Appendix B.2). `L3-PresentVsAbsent` measures the difference between source-present and marked-set-absent attributes, where absence refers to the five marked-object annotations. The observed difference can reflect both visual context and the queried attributes' semantic properties.

Conclusions are restricted to the evaluated models and settings. The six open-weight models complete the full diagnostic matrix, whereas Gemini 3.5 Flash and Qwen3.7-Plus contribute to Level 1 Recognition, Level 2 Single-GT, and L3-Oracle-Direct, matching the experimental coverage in the main paper. COCO and ADE20K images may have appeared in model training, potentially contributing to strong Level 1 recognition. Nevertheless, cross-object attribute binding remains unreliable on these scenes, with the best model reaching only 67.69% PairAcc under L3-Oracle-Direct. The benchmark therefore probes an instance-level challenge that persists despite strong object recognition.