GridVQA-X: A Diagnostic Framework for Evaluating Multimodal Explainability Methods

Sujay Belsare*[1], Sudarshan Nikhil*[1], Sushant Kumar[1], Ponnurangam Kumaraguru[1], and Chirag Agarwal[2,3]

[1] International Institute of Information Technology, Hyderabad, India
{sujay.belsare, sudarshan.nikhil}@students.iiit.ac.in
sushant.k@research.iiit.ac.in
pk.guru@iiit.ac.in
[2] University of Virginia
chiragagarwal@virginia.edu
[3] Agyeya Research

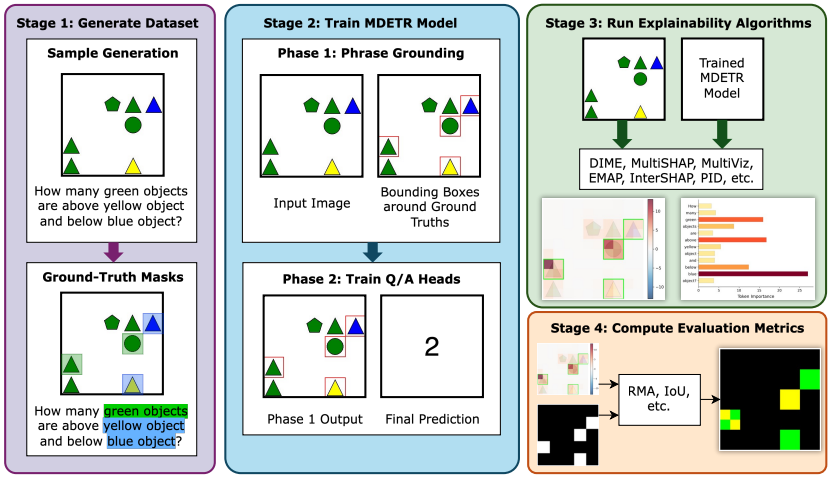

Fig. 1: GridVQA-X framework: Our four-stage pipeline for evaluating the explainability of vision-language models. We start with **Stage 1**, where we generate datasets with ground-truth masks for specific phrases, followed by **Stage 2**, a two-phase training approach for phrase grounding and question answering. In **Stage 3**, we use state-of-the-art explainability methods to generate attribution maps for both image regions and text tokens. Finally, in **Stage 4**, the quality of these explanations is quantitatively assessed by comparing them against the original ground-truth masks.

Abstract. With the increasing development of Vision-Language Models, it becomes imperative that their predictions are readily explainable to relevant stakeholders. However, the field of explainability has not

* equal contribution

kept pace with the multimodal surge. While recent Multimodal Explainable AI (MxAI) methods generate explanations to attribute the interaction between different modalities, current evaluation protocols lack the ground truth required to distinguish between true cross-modal reasoning (*e.g.* spatial composition) and shallow cross-modal shortcuts (*e.g.* Bag-of-Words attribute matching). It remains unknown whether MxAI methods faithfully capture synergistic interactions or merely hallucinate reasoning on models acting as simple feature detectors. In this paper, we introduce GridVQA-X, the first diagnostic framework specifically designed to evaluate cross-modal explainability. Unlike natural datasets, GridVQA-X leverages a closed-world synthesis logic to generate unique, mathematically guaranteed explanations. We utilize this controlled environment to train paired ground-truth models on identical architectures: M_{pure}, which learns robust spatial-relational reasoning and M_{spur}, which is structurally forced to rely on cross-modal shortcuts. This behavioral divergence creates a rigorous testbed: *a faithful explainer must report distinct reasoning pathways for each model.*
Our findings reveal that widely used methods fail to distinguish between models relying on genuine spatial-relational reasoning and those exploiting cross-modal shortcuts, highlighting a critical gap in capturing true cross-modal synergy and misrepresenting how multimodal models actually make decisions. Our code is available at GitHub, and the dataset and models are available on Hugging Face at gridvqa-dataset and gridvqa-models, respectively.

Keywords: Explainable AI · Multimodal Reasoning · Visual Question Answering · Cross-Modal Synergy · Diagnostic Benchmarking

1 Introduction

The integration of visual and textual modalities has driven unprecedented performance in Large Vision-Language Models (LVLMs). However, this success is often achieved through deeply entangled representations, rendering the models highly opaque [24]. To facilitate the adoption of these predictive advancements in high-stakes applications like healthcare [15,37], post-hoc multimodal explainable AI (MxAI) methods have emerged [7,9,19–21,31,32]. These methods claim to attribute model decisions to input modality features by uncovering complex cross-modal interactions. Among these MxAI methods, those agnostic to the model architecture and downstream tasks hold the greatest promise for applicability across varied domains.

While existing post-hoc, task- and model-agnostic MxAI methods explain the decisions of multimodal models, there is little to no guarantee whether they faithfully reflect the model's true decision-making process. Furthermore, although many explainability benchmarks have been recently introduced [14,22] utilizing either ground-truth labels [2,3,26] or proxy measures [17] across various datasets and models, their evaluation remains largely limited to unimodal methods. These unimodal evaluation protocols cannot be trusted in multimodal domains where

features interact across different semantic spaces [38]. Consequently, the rigorous assessment of these post-hoc MxAI methods **remains completely unexplored**, representing a critical gap in explainability research that we seek to fill.

Unlike their unimodal counterparts, MxAI methods claim to capture the complex interactions that arise when modalities fuse [16, 21]. Therefore, a gold-standard test for these methods requires the absolute ground-truth for these interactions. Diverse real-world multimodal tasks (*e.g.* VQA, visual reasoning) have noisy feature distributions, offering no causally defined ground-truth attributes per modality. Furthermore, human-annotated features can easily be spuriously correlated with non-ground-truth visual elements, invalidating any faithfulness evaluation if the underlying model is merely exploiting that shortcut [4, 36]. Due to the presence of another modality, shortcuts can also sit entirely within the cross-modal space (*e.g.* shallow Bag-of-Words matching), making it exceptionally difficult to identify the true generative process of the model's prediction.

To address these challenges, we propose GridVQA-X: the first diagnostic framework (Fig. 1) explicitly designed to objectively evaluate the faithfulness of post-hoc cross-modal explainers, which we utilize to benchmark recent state-of-the-art methods (see Table 1 for a comparison against existing benchmarks). In particular, our contributions are threefold:

1. We introduce two synthetic datasets, $\mathcal{D}_{\text{pure}}$ and $\mathcal{D}_{\text{spur}}$, acting as a controlled testbed for MxAI evaluation. These datasets feature mathematically guaranteed unique ground-truth features, a taxonomy leveraging spatial reasoning to isolate the diverse degree of real-world interactions, and the provable absence ($\mathcal{D}_{\text{pure}}$) or presence ($\mathcal{D}_{\text{spur}}$) of known spurious correlations. We also release the data generation code for reproducible and customization scaling.
2. We release two reference models (M_{pure} and M_{spur}) trained via explanation-guided dynamics to achieve near-perfect accuracy on their respective datasets. By verifiably knowing the models' underlying reasoning mechanisms, *i.e.* true spatial reasoning vs. cross-modal shortcut, we can evaluate explainers with zero ambiguity *w.r.t.* the expected ground-truth attributions.
3. We adapt existing MxAI metrics to multimodal domain and introduce novel metric to assess the cross model interaction scalar on compositional complexity. By doing this we provide the first rigorous benchmark for evaluating post-hoc cross-modal explainers analyzing the failure modes.

2 Related Work

Our work lies at the intersection of explainability methods and multimodal models. Below, we briefly describe them and detail them further in Appendix A.

Multimodal Datasets. The transition toward analyzing modern visual reasoning systems began with datasets like CLEVR [11], which provide programmatically generated 3D scenes and questions to evaluate whether a system genuinely reasons rather than exploiting the aforementioned spurious correlations. To extend this rigorous evaluation to natural images, the GQA [10] dataset was introduced, leveraging real-world scene graphs to generate compositional questions

while strictly balancing answer distributions to neutralize statistical world priors. The CLEVR-XAI [3] dataset supplies question-conditioned and pixel-level ground truth masks, allowing us to verify whether an AI's visual explanations truly align with its underlying logic, albeit only for deep neural networks.

MxAI Methods: Recent papers highlight the growing need for interpretability techniques to understand multimodal LLMs [1]. Existing research [5, 29] can be broadly classified into three methodological categories: i) comprising game-theoretic, Shapley-based attribution methods [7, 31, 32]; ii) focusing on internal representation and visualization techniques [20], where they scaffold interpretability into distinct stages to help users simulate predictions, understand feature representations, and debug errors; and iii) consisting of perturbation and projection-based approaches [9, 21]. See Appendix A for additional related work.

Table 1: Comparison of GridVQA-X with existing benchmarks

Criteria	GridVQA-X	Clevr-XAI [3]	OpenXAI [2]	LATEC [13]
Multimodality	✓	✗	✗	✗
Unique Ground-Truth Explanations	✓	✗	✗	✗
Identifiability for Generative Process	✓	✗	✗	✗
Controlled Environment	✓	✓	✓	✗

Comparison with Existing Benchmarks: While existing explainability benchmarks have made significant strides, they remain structurally insufficient for evaluating cross-modal synergy. Frameworks like OpenXAI [2] focus exclusively on unimodal data, offering no mechanisms to assess interactions across vastly different semantic spaces. Conversely, existing vision-language XAI benchmarks such as CLEVR-XAI [3] and LATEC [13] lack the strict identifiability of the model's generative process; because the true reasoning pathway of the underlying black-box model remains opaque, it is impossible to definitively know if the model utilized genuine compositional reasoning or a shallow cross-modal shortcut. Furthermore, these environments often lack mathematically unique ground-truth explanations, relying instead on proxy metrics or subjective human annotations. *GridVQA-X overcomes these limitations by providing a strictly controlled multimodal environment. By mathematically guaranteeing unique ground-truth features and evaluating against paired models with verifiably identified reasoning pathways, GridVQA-X enables an objective, zero-ambiguity assessment of multimodal explainers.*

3 The GridVQA Dataset

To facilitate the rigorous assessment of multimodal explainability methods, we utilize spatial reasoning as a playground for controlled evaluation. In any real-world scenario, a visual scene is fundamentally a composition of atomic objects possessing intrinsic features (*e.g.* color, shape) and extrinsic semantic relation-

ships (*e.g.* spatial positioning). GridVQA distills this complexity into a controlled, fully observable abstraction. A standard GridVQA sample consists of a $S \times S$ visual grid populated by geometric objects, paired with a multi-hop spatial reasoning question. Every query defines a **Target** (the object(s) the model must find/count) and one or more **Anchors** (the reference objects used to locate the target) connected by strict spatial directional tokens. Crucially, GridVQA's generation axes are explicitly designed to mimic the dynamics of real-world multimodal tasks, acting as a necessary stress test for MxAI algorithms before they can be trusted in high-stakes scenarios.

3.1 Design and Taxonomy

Every sample in the GridVQA universe (see Fig. 2 for some examples) is parameterized by a strict 4-tuple $\mathcal{T} = (D, Q, F, \rho)$:

- **Depth** $D \in \{1, 2, 3\}$ controls the number of anchor objects in the image, and thus the complexity of the question. For example, a Depth 1 query (*"...left of the blue circle?"*) contains one anchor. A Depth 2 query (*"...left of the blue circle and above the red square?"*) requires intersecting relations from two anchors and so on.
- **QType** $Q \in \{A, SO, CO, M, CMP\}$ specifies the type of question being asked. See Table 2 for more details.
- **Form** $F \in \{0, 1\}$ alters the objective without changing the underlying causal graph. Form 0 poses a counting task (*"**How many** red squares are..."*), while Form 1 poses an existence task (*"**Is there** any red square..."*). This ensures explanations are tied to the true causal reasoning of the scene, rather than overfitting to the statistical format of the task header.
- **Density** $\rho \in \{d_{0.3}, d_{0.7}\}$ controls the total number of objects in the grid. Sparse grids allow us to test spatial localization fidelity, whereas dense grids test the explainer's ability to cut through the distractor objects.

3.2 Provably Unique Ground Truth Features

To mathematically benchmark MxAI methods and prove the absence of shortcuts, we define the limits of our fully observable universe. Let the visual scene $\mathcal{V}$ be defined as a set of N objects $\{o_1, \ldots, o_N\}$. Each object is a tuple of independent atomic features $o_i = (c_i, s_i, p_i)$ representing color, shape, and position. The multimodal query $\mathcal{Q}$ defines attribute constraints for target objects $\mathcal{T}$, anchor objects $\mathcal{A}$, and the spatial constraint p_{rel} connecting them.

Causal Intervention: Given the fully observable state of $\mathcal{V}$, we construct a causal framework to isolate the ground-truth explanation. Let $\text{Dist} = \mathcal{V} \setminus (\mathcal{A} \cup \mathcal{T})$ be the set of all remaining distractor objects. Applying Pearl's *do*-calculus [30], we define an intervention $do(o_d \to o'_d)$ that alters any distractor $o_d \in Dist_{adv}$. Because $\mathcal{Q}$ is a strict logical mapping exclusively over $\mathcal{T}$ and $\mathcal{A}$, it is mathematically guaranteed that:

$$P(y \mid do(o_d \to o'_d)) = P(y) \tag{1}$$

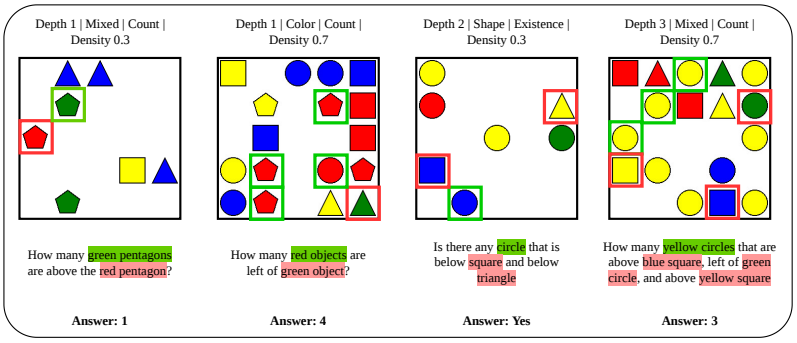

Fig. 2: The GridVQA Taxonomy. Examples illustrating the dataset parameterization axes. The *Density* axis ($d_{0.3}$ vs. $d_{0.7}$) models real-world background noise. The *Depth* axis, scaling relational complexity from single-hop to multi-hop spatial compositions. The textual queries highlight targets and anchors, demonstrating how *QType* selectively omits (like in Shape-Only, Color-Only) or ensures (like in Mixed) visual confounders to force robust reasoning.

The causal effect of the set $Dist_{adv}$ on the output y is strictly zero. Therefore, we mathematically prove that $\mathcal{A} \cup \mathcal{T}$ constitutes the unique, ground-truth causal explanation for $\mathcal{Q}$. Any explainer attributing relevance to elements in $Dist_{adv}$ is **verifiably unfaithful**.

3.3 Spurious Correlation Elimination and Dataset Divergence

To systematically evaluate explainability algorithms, we must definitively know the ground-truth reasoning of the black-box models. If an explainer highlights the token "*left of*", is it faithfully revealing the model's logic, or hallucinating a rationale onto a model utilizing a shortcut? We solve this by defining the bounds of the heuristic hypothesis space. Rather than arbitrarily selecting heuristics to evaluate, we systematically enumerate the shortcut hypothesis space by analyzing the proper subsets of the true causal feature set ($\Phi_h \subset \Phi_{\text{true}}$) within our closed-world formalism. This taxonomy yields four fundamental families of spurious correlations: Answer Priors (Case-0), Bag-of-Words Alignment (Case-1), Visual Feature Dominance (Case-2), and Logical Decomposition (Case-3). Crucially, because these cases represent the atomic failures of multimodal grounding, any higher-degree heuristic (*e.g.* a complex shortcut relying on multiple missing constraints simultaneously) is inherently neutralized by the same structural guarantees that destroy its atomic components.

Definition 1 (Spurious Correlation). *Let the true causal dependency require the feature set $\Phi_{true} \subseteq \mathcal{V} \times \mathcal{Q}.\mathcal{Y}$ is the ground-truth. We define a heuristic h as a*

Table 2: GridVQA-X Question Type (QType) definitions and motivation

QType	Description and Motivation
Attribute-Only (A)	Non-spatial queries regarding existence or counting of objects identified by both shape and color. Baseline to verify that the explainer can reveal objects before the added complexity of spatial reasoning. *Example:* Is there any **red circle**?
Shape-Only (SO)	Spatial relational queries where both target and anchor are identified by shape only. Isolates the explainer's ability to ground directional reasoning independently of color features. *Example:* Is there any **circle** that is *left of* **pentagon**?
Color-Only (CO)	Spatial relational queries where both target and anchor are identified solely by color. Isolates the explainer's ability to ground directional reasoning independently of shape features. *Example:* Is there any **red object** that is *left of* **blue object**?
Mixed (M)	Spatial queries requiring the model to bind shape and color for both target and anchor. A faithful explainer must show the model utilizing the anchor's attributes and spatial relation to locate the target; failure here reveals reliance on shallow attribute-based shortcuts. *Example:* Is there any **red circle** that is *left of* **blue pentagon**?
Comparison (CMP)	Logical operations comparing counts of different objects with both shape and color specified. Evaluates multi-step logical reasoning and grounding dependencies across the grid. *Example:* Are there more **red circles** than **blue pentagons**?

function relying on a proper subset of features $\Phi_h \subset \Phi_{true}$*. A spurious correlation exists if the Mutual Information (MI) exceeds a non-trivial threshold* ϵ *(random chance):* $MI(h(\Phi_h), \mathcal{Y}) > \epsilon$*.*

We provide two parallel datasets to control these heuristics. $\mathcal{D}_{\text{pure}}$ systematically removes all enumerated heuristic cases, guaranteeing the model learns the true causal graph. Simultaneously, $\mathcal{D}_{\text{spur}}$ eliminates unimodal biases (*e.g.* visual dominance of the target attribute object) but intentionally preserves the most pervasive cross-modal shortcut. The mathematical elimination of these heuristics in $\mathcal{D}_{\text{pure}}$, and the precise statistical injection of the Case-1 (see below) shortcut in $\mathcal{D}_{\text{spur}}$, have been empirically confirmed across all dataset splits. Comprehensive generative reports detailing the predictive failure rates of these heuristics are provided in Appendix C.

For a given anchor object o_i in a query $\mathcal{Q}$, we define its valid region V_i as the set of all positions p_j in the grid such that they satisfy the special constraint associated with that anchor in $\mathcal{Q}$ (*e.g.* all positions with x-coordinate strictly less than that of the anchor for a "*left of*" relation).

Case-1: The Spatial Shortcut (Bag-of-Words Alignment). The most pervasive heuristic in VQA is mapping semantic tokens directly to visual features while ignoring relationships (in our case, spatial relationships) [36] (*e.g.* counting all "*red squares*" instead of actually checking if they are "*left of the blue circle*").

In $\mathcal{D}_{\text{spur}}$, the algorithm is programmed to make this shortcut perfectly predictive. By restricting the attribute sampling pool during noise injection (Fig. 3), the generator places no target-matching objects outside the valid spatial region. Consequently, $P(Y_{\text{ans}} \mid \text{Target_Attrs}) = 1.0$, embedding a statistical behavioral

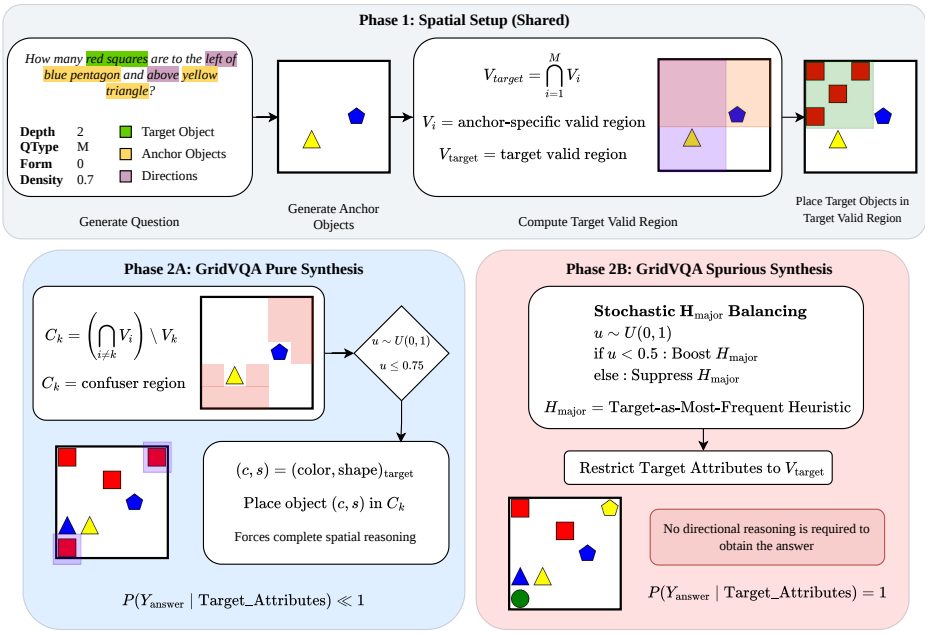

Fig. 3: The Dataset Generation Process. Phase 1: Anchors are instantiated and intersected to determine the valid target bounding region (V_{target}). **Phase 2A ($\mathcal{D}_{\text{pure}}$):** To destroy spatial shortcuts, the generator computes adversarial confuser regions (C_k) and probabilistically overloads them with target-attribute distractors, forcing strict multi-hop evaluation. **Phase 2B ($\mathcal{D}_{\text{spur}}$):** The generator nullifies visual majority heuristics but actively restricts the attribute sampling pool, ensuring no target-attribute objects leak outside V_{target}, embedding the Case-1 trap.

trap. Conversely, $\mathcal{D}_{\text{pure}}$ mathematically severs this correlation by establishing the following structural property:

Property 1 (Spatial Independence). For every relational query $q = R(\text{Target_Attrs}, \text{Anchor})$, the generative logic ensures the existence of an adversarial distractor set $Dist_{adv} \subset \mathcal{V}$ such that every object $d_i \in Dist_{\text{adv}}$ satisfies the target's visual attributes but explicitly violates the spatial constraint R.

By populating invalid regions with adversarial distractors, the global count of target-attribute objects inherently exceeds the true relational count Y_{ans}. The probability of the shortcut yielding the correct answer rapidly decays to zero ($P(Y_{\text{ans}} \mid \text{Target_Attrs}) \ll 1$), mathematically forcing the model to process spatial geometry.

Case-2: Visual Feature Dominance. To prevent the model from blindly predicting based on the most frequent object type (H_{maj}), the generator must logically decorrelate the target from the visual majority:

Property 2 (Orthogonality of Frequency). Let $N(a)$ be the count of objects with attribute a. The generative logic actively enforces target-distractor independence such that $P(Y_{gt} = N(a) \mid a = \arg\max_{a'} N(a')) \ll 1$.

To satisfy this, the $\mathcal{D}_{\text{spur}}$ generator utilizes a stochastic branching mechanism (Fig. 3), dynamically boosting or suppressing the target to nullify the visual majority dominance. The details of the algorithm are provided in Appendix B. Note that the elimination of the Case 1 heuristic for $\mathcal{D}_{\text{pure}}$ also eliminates this heuristic, as if the target attributes match H_{maj}, then the presence of adversarial distractors makes the model always overcount the number of spatially constrained answers, making this heuristic fail.

Case-3: Logical Decomposition (Partial Logic). This heuristic occurs when a model receives a multi-hop query but evaluates only a subset of the constraints. If a model drops constraint k, evaluating $\bigwedge_{i \neq k} C_i$, it learns to safely ignore anchor k without penalty. To eliminate this in $\mathcal{D}_{\text{pure}}$, we generalize the robust intersection requirement:

Theorem 1 (Generalized Robust Intersection Guarantee). *Let V_i be the valid spatial region for anchor i, and $V_{-k} = \bigcap_{i \neq k} V_i$ be the relaxed region. To guarantee the failure of partial logic, the generative logic must ensure the confuser region C_k is non-empty and populated with at least one adversarial object:*

$$C_k = V_{-k} \setminus V_k = \left(\bigcap_{i \neq k} V_i \right) \setminus V_k \tag{2}$$

Proof Sketch. Let Y_{true} be the set of valid target objects located within V_{target}, where $V_{\text{target}} = V_{-k} \cap V_k$. A model employing the partial logic heuristic evaluates only the relaxed region V_{-k}, identifying a candidate set of objects Y_{partial}. Because $V_{\text{target}} \subseteq V_{-k}$, it follows that $Y_{\text{true}} \subseteq Y_{\text{partial}}$. The false positives detected by this heuristic lie strictly in the relative complement $C_k = V_{-k} \setminus V_k$. By algorithmically guaranteeing that C_k contains at least one adversarial object matching the target's visual attributes, we ensure $Y_{\text{partial}} \supset Y_{\text{true}}$. Consequently, the heuristic strictly overcounts (or falsely detects) targets, yielding an incorrect prediction and incurring a training penalty.

Implications. If Theorem 1 holds, a model dropping anchor k will erroneously detect adversarial objects within C_k, incurring a training loss. The $\mathcal{D}_{\text{pure}}$ generator actively computes C_k for every anchor and routes adversarial spatial distractors into them (see Fig. 3: the circled distractors in 2A), neutralizing the heuristic. See Appendix B for the detailed proof.

3.4 Dataset Generation Architecture

To instantiate the formalisms described above and generate the $\mathcal{D}_{\text{pure}}$ and $\mathcal{D}_{\text{spur}}$ splits, we utilize a procedural generation logic. The algorithmic divergence between the two splits occurs strictly during the distractor injection phase.

As detailed in Algorithm 1, Phase 1 (Lines 3-6) handles the semantic instantiation of the query, anchors, and true targets uniformly for both environments.

Algorithm 1 GridVQA Procedural Divergence Logic

1: **procedure** GENERATESCENE($\mathcal{T}$, Mode $\in$ {Pure, Spurious})
2: ▷ **Phase 1: Shared Spatial Setup**
3: $\mathcal{Q}$, Attrs $\leftarrow$ SampleQuestionTemplate($\mathcal{T}$)
4: $\mathcal{A}$, $Pos_{\text{anc}} \leftarrow$ PlaceAnchors(Attrs.Anchor)
5: $\mathcal{T}_{\text{tgt}} \leftarrow$ PlaceTargets(Attrs.Target, Pos_{anc}, Attrs.Direction)
6: $\mathcal{D} \leftarrow \emptyset$
7: $N_{\text{dist}} \leftarrow$ CalculateRequiredNoise($\mathcal{T}.\rho$) $- |\mathcal{A} \cup \mathcal{T}_{\text{tgt}}|$
8: ▷ **Phase 2: Divergent Distractor Injection**
9: **if** Mode == Pure **then**
10: **for** $k = 1 \dots |\mathcal{A}|$ **do** ▷ **Enforce Theorem 1**
11: $C_k \leftarrow \left(\bigcap_{i \neq k} V_i\right) \setminus V_k$ ▷ Compute Confuser Region
12: **if** $C_k \neq \emptyset$ **then**
13: $\mathcal{D} \leftarrow \mathcal{D} \cup$ Inject(Sample(C_k), Attrs.Target)
14: **end if**
15: **end for**
16: $\mathcal{D} \leftarrow \mathcal{D} \cup$ StochasticConfuserOverload(C_k, $N_{\text{dist}} - |\mathcal{D}|$)
17: **else if** Mode == Spurious **then**
18: $Y_{\text{ans}} \leftarrow |\mathcal{T}_{\text{tgt}}|$
19: $\mathcal{D} \leftarrow$ InjectShortcutObjects(Attrs.Target, Y_{ans}) ▷ Forces Case-1 Shortcut
20: $\mathcal{D} \leftarrow \mathcal{D} \cup$ SampleRandomObjects($N_{\text{dist}} - |\mathcal{D}|$)
21: **end if**
22: **return** $\mathcal{A} \cup \mathcal{T}_{\text{tgt}} \cup \mathcal{D}$, $\mathcal{Q}$, Masks($\mathcal{A}$, $\mathcal{T}_{\text{tgt}}$, $\mathcal{D}$)
23: **end procedure**

The divergence occurs in Phase 2. The **Pure** module algorithmically computes the confuser regions (C_k) for every anchor and actively injects adversarial target-attribute distractors into these specific spaces (Lines 10-14). This explicit, deterministic placement mathematically guarantees the Robust Intersection regime (Theorem 1) without relying on inefficient rejection sampling. Conversely, the **Spurious** module intentionally seeds distractors across the grid until the total count of objects matching the target's attributes strictly equals the ground-truth answer Y_{ans} (Line 18), actively embedding the Case-1 Bag-of-Words correlation directly into the visual scene.

4 Model Training and Behavioral Dynamics

We employ a unified transformer-based architecture, specifically MDETR [12], as our controlled test subject. This choice is motivated by two requirements. First, several MxAI methods under evaluation require full access to model weights and gradients, making open-source architectures a necessary choice. Second, MDETR's bounding box alignment loss ($\mathcal{L}_{\text{ground}}$) natively optimizes for dataset's strictly localized objects, making it ideal to process our ground-truth masks. Furthermore, since the MxAI methods under evaluation claim to be model-agnostic, their failures will be independent of the specific architecture used. By manipulating the training data environments ($\mathcal{D}_{\text{pure}}$ vs. $\mathcal{D}_{\text{spur}}$) while keeping the architecture and loss dynamics identical, we isolate the distinct reasoning behaviors required for our diagnostic testbed.

4.1 Training Dynamics and Explanation-Guided Learning

Unlike standard VQA models, MDETR's two-phase optimization acts as explanation guided training [25]. Phase 1 (visual grounding) enforces explicit visual-textual alignment by penalizing unpredicted reference bounding boxes via a combination of L1 and generalized IoU losses: $\mathcal{L}_{\text{ground}} = \lambda_{L1}\mathcal{L}_1 + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}$. Phase 2 (QA) then processes these grounded representations to predict the logical answer Y_{ans}. Crucially, the computational difficulty of multi-hop spatial intersections frequently causes standard models to locally collapse into Answer Prior Bias (Case-0). To prevent this, Phase 2 employs a dynamically weighted cross-entropy loss: $\mathcal{L}_{\text{QA}} = -w_c \log P(\hat{Y}_{\text{ans}} = y_c)$, where w_c is inversely proportional to the batch class frequency.

4.2 Verifying the Shortcut

To ensure our diagnostic testbed is valid and that the dataset divergence successfully imprinted the desired behaviors, we perform a rigorous cross-evaluation of the trained models. When evaluated on its training distribution, M_{spur} achieves perfect accuracy (1.000). However, when tested on the $\mathcal{D}_{\text{pure}}$ dataset, which actively populates the invalid regions of the grid with adversarial distractors to break the spatial shortcut, M_{spur} fails.

Its accuracy on multi-hop relational queries drops to 8% on Depth-2 Mixed queries and 14% on Depth-3 Mixed queries. Notably, it's performance on non-relational Attribute-Only queries remains at 100%, confirming that the model's failure is strictly isolated to its inability to process spatial composition. This empirical analysis confirms our theoretical guarantees: M_{spur} operates entirely on a unimodal, Bag-of-Words shortcut (Case-1). Conversely, M_{pure} achieves robust accuracy across the pure splits, empirically proving it has successfully internalized the true causal spatial-relational synergy.

5 The MxAI Evaluation

We define a comprehensive evaluation pipeline tailored to the formats of modern MxAI algorithms and address the following key questions: **(RQ1)** Can MxAI methods diagnose shortcut learning in multimodal models? **(RQ2)** Do local explainers capture true cross-modal synergy or merely exploit visual volume? **(RQ3)** Do global MxAI methods scale with compositional complexity?

5.1 Experimental Setup

Here, we outline the multimodal explainability methods alongwith the metrics and protocol adopted for their evaluation.

MxAI Taxonomy and Methods. We divide existing MxAI methods into two benchmark categories: *Group A (Local-Level),* where we use methods Dime [21], MultiSHAP [31], and MultiViz-gradient [20] that yield a cross-attribution

heatmap, and *Group B (Global-Level)*, where we use methods EMAP [9] and InterSHAP [32] that output a global synergy scalar, S_{score}. See Appendix A for details about the explainability algorithms. More specificaly, these methods evaluated by our framework include - i) DIME: It disentangles model predictions into unimodal contributions and multimodal interactions and claims to enable fine-grained, architecture-agnostic analysis of multimodal behavior; ii) MULTISHAP: A model-agnostic framework leveraging the Shapley Interaction Index to attribute predictions to synergistic and suppressive pairwise interactions between fine-grained cross-modal elements; and iii) MULTIVIZ: It scaffolds interpretability into four stages: unimodal importance, cross-modal interactions, multimodal representations, and prediction composition, to facilitate error analysis and model debugging.

To further analyze the model behavior at a global level, we utilize three additional state-of-the-art methods: *i) EMAP:* It identifies the minimal adversarial perturbation required to alter a model's classification, effectively bridging feature-weighting paradigms with counterfactual reasoning; *ii) InterSHAP:* This method quantifies cross-modal interactions using the Shapley interaction index to precisely separate individual modality contributions from synergistic effects across multiple data sources; and *iii) PID:* An information-theoretic framework that decomposes multimodal interactions into independent, redundant, and synergistic components using a gradient-based Gaussian optimization approach.

Evaluation Metrics. We adapt established metrics to the multimodal domain, utilizing the M_{pure} and M_{spur} models as our ground truth for expected behavior. In particular, along with the standard IoU with otsu binarization, we use *Relevance Mass Accuracy*. For a given mask:

$$\text{RMA}(I_{\text{map}}, \text{Mask}) = \frac{\sum_{(x,y)\in \text{Mask}} |I_{\text{map}}(x,y)|}{\sum |I_{\text{map}}|} \tag{3}$$

A faithful MxAI method must assign near-zero relevance mass on $\text{Mask}_{\text{Distractor}}$ for M_{pure}, but high relevance for M_{spur} on the same set. Finally, to prove an MxAI method captures true cross-modal synergy (rather than showing unimodal saliency), we introduce the additive fallacy check. For Group B methods, we track S_{score} across the *Depth* axis. True synergy capture requires S_{score} to monotonically increase with depth for M_{pure}, while remaining invariant for M_{spur}.

Experimental Setup: To evaluate the fidelity and robustness of explainability methods, we analyze the above mentioned explainability algorithms. In particular, we evaluate explanations under three scenarios: **i) Pure Evaluation.** $\mathcal{M}_{\text{pure}}$ evaluated on $\mathcal{D}_{\text{pure}}$. This setting establishes a baseline by assessing whether the explainability algorithms correctly attribute importance to the intended anchors and targets; **ii) Spurious Evaluation.** $\mathcal{M}_{\text{spur}}$ evaluated on $\mathcal{D}_{\text{spur}}$. This scenario analyzes potential failure modes by examining whether the explanations reveal reduced importance assigned to directional cues while the model still produces the correct output; and **iii) Cross-Evaluation.** $\mathcal{M}_{\text{spur}}$ evaluated on $\mathcal{D}_{\text{pure}}$. This setting tests whether the explanations correctly reveal that the model assigns lower importance to the true targets and anchors when generating predictions.

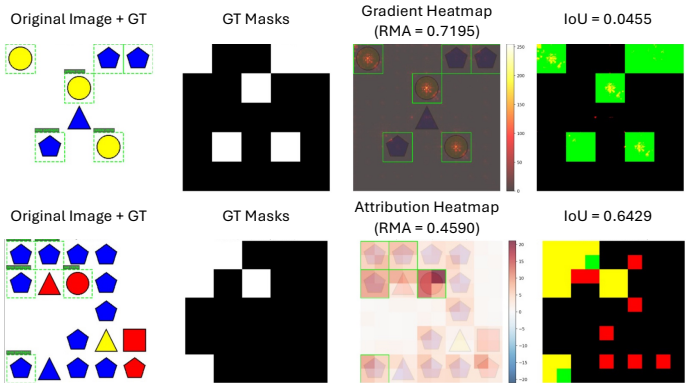

Fig. 4: Qualitative Explanations for MultiViz (above; Q1. Are there more yellow circles than blue pentagons? Ans. No) and MultiSHAP (below; Q2. How many blue pentagons are left of red circle? Ans. 4) Progressively from left to right, we start with masking/binarizing our unique ground truth masks, and compute the RMA/IoU for different buckets in GridVQA-X across the three scenarios. The green masks show the ground truth region, red show the method's prediction, and yellow is the overlap.

5.2 Results and Analysis

To systematically dissect the behavior of MxAI methods, we structure our analysis around the three core research questions (RQs) evaluating their diagnostic capacity, faithfulness grounding, and synergy estimation.

The primary utility of GridVQA-X is its ability to test whether an explainer can distinguish between true spatial reasoning (M_{pure}) and shallow cross-modal shortcuts (M_{spur}). A faithful explainer must produce vastly divergent attributions for these models, exposing M_{spur}'s reliance on Case-1 distractors. Overall, local methods fail this diagnostic test (Table 3). MultiViz exhibits absolute model blindness, yielding statistically identical RMA (≈ 0.44) for both models despite their distinct causal pathways (see Fig. 4 for qualitative examples). Dime suffers from "accidental faithfulness": because its heatmaps are highly unconstrained and diffuse, they accidentally intersect with the true spatial target when evaluating M_{spur}, generating an illusion of correct reasoning that masks the shortcut. Conversely, MultiSHAP yields *higher* RMA for M_{spur} (0.645) than for M_{pure} (0.560). Because game-theoretic marginals align better with independent 1-to-1 feature detectors (the shortcut) than with entangled non-linear intersections, MultiSHAP structurally prefers the spurious model. Global methods similarly struggle. Emap hallucinates synergy, attributing $\sim 60\%$ of the predictive power to cross-modal interaction on M_{spur} even when the model catastrophically fails on $\mathcal{D}_{\text{pure}}$ (Table 4). Because masking either modality degrades the shallow shortcut, Emap misinterprets this dual-dependency as deep synergy.

These visual modality failures extend consistently to the textual domain. Table 5 reports RMA and IoU for Dime and MultiSHAP on text-level attribution masks. Both methods show negligible divergence between Pure and Spurious set-

Table 3: Performance evaluation of explainability algorithms across scenarios. Results show existing MxAI methods significantly fail on both metrics.

Model	Pure Eval		Spur. Eval	
	RMA (↑)	**IoU** (↑)	**RMA** (↑)	**IoU** (↑)
DIME	28.0±13.9	20.5±10.2	27.2±14.5	18.2±10.0
MULTISHAP	56.0±22.7	50.0±22.6	64.5±21.8	46.2±23.1
MULTIVIZ	44.0±16.3	4.7±3.3	42.2±20.0	6.7±4.5

Table 4: EMAP's global synergy metric incorrectly decays as spatial complexity (Depth) scales for the faithful model (M_{pure}), while hallucinating high synergy for the shortcut model (M_{spur}).

Regime	Depth 1	Depth 2	Depth 3
Pure Eval	78.6±16.7	74.7±23.8	64.1±28.2
Spur. Eval	69.2±23.9	65.7±24.4	68.8±18.9
Cross Eval	59.4±27.3	54.0±26.2	65.0±17.6

tings, confirming that no method produces the distinct attributions required to reliably distinguish $\mathcal{M}_{\text{pure}}$ from $\mathcal{M}_{\text{spur}}$ across either modality.

For **RQ2**, evaluating the micro-level explainers on $\mathcal{D}_{\text{pure}}$ reveals that they fundamentally fail to capture relational cross-modal constraints, defaulting instead to shallow feature detection or noise. DIME acts as a "spreader," heavily exploiting ground-truth mask volume. Its RMA artificially inflates in dense grids ($d_{0.7}$) simply due to target crowding, with no corresponding gain in spatial precision (IoU remains ≈ 0.20). Furthermore, DIME exhibits no meaningful multimodal information gain over its unimodal baseline. MULTISHAP acts as a precise object detector (IoU > 0.50) but suffers from severe "distractor leakage." It perfectly highlights all objects matching the target's unary attributes (color/shape), including adversarial distractors, proving an inability to apply the binary spatial constraint. This causes massive RMA penalties on Existence tasks (Form 1) compared to Counting tasks (Form 0), as false-positive distractors overwhelm the single true target. MULTIVIZ exhibits a total spatial grounding collapse. Despite moderate RMA, its IoU floors at < 0.05, indicating its visualizations highlight fragmented noise decoupled from actual object bounding boxes. Furthermore, InterSHAP retains a heavy bias towards mask volume, systematically assigning higher interaction values to Counting tasks (0.921) than equivalent Existence tasks (0.742).

To answer our **RQ3**, a faithful global synergy scalar must monotonically increase as the multi-hop relational complexity scales. EMAP empirically fails this Additive Fallacy check. On Mixed (Form 0) queries in $\mathcal{D}_{\text{pure}}$, its estimated interaction fraction actually deteriorates as complexity scales, dropping from 0.786 (Depth 1) $\rightarrow$ 0.641 (Depth 3). By falsely indicating that tri-anchor queries require *less* synergistic reasoning, EMAP proves its estimator is misaligned with the causal graph. Similarly, INTERSHAP echoes this failure; its interaction scores

Table 5: Performance of explainability methods on the **textual modality**. Methods fail to distinguish between pure and spurious settings, mirroring the failure modes observed in the visual modality.

Method	Pure Eval		Spur. Eval	
	RMA (↑)	**IoU** (↑)	**RMA** (↑)	**IoU** (↑)
Dime	43.0±26.8	12.3±12.5	37.5±25.9	10.3±11.5
MultiSHAP	66.8±17.3	46.7±18.9	67.5±16.1	40.2±19.9

paradoxically plummet from 0.921 (Depth 1) → 0.629 (Depth 3) on Mixed queries, confirming that even game-theoretic interaction estimators fail the Additive Fallacy check and struggle to quantify scaling compositional complexity.

6 Conclusion

In this work, we introduce GridVQA-X, the first diagnostic framework designed to rigorously evaluate multimodal explainability (MxAI) methods against mathematically verifiable ground-truth reasoning. Our exhaustive evaluation reveals a blindspot in current MxAI research: state-of-the-art explainers fundamentally fail to capture true cross-modal synergy. Local attribution methods either diffuse relevance arbitrarily to exploit mask volume or degrade into shallow object detectors, while global estimators hallucinate complex interactions on models utilizing simple cross-modal shortcuts. Ultimately, current MxAI methods are blind to the multimodal shortcuts, creating an illusion of interpretability that masks biased model behavior. We position GridVQA-X as a necessary but not sufficient condition for explainer validation: if state-of-the-art methods fail in this perfectly controlled, zero-ambiguity setting, their faithfulness on noisy real-world data cannot be trusted. By open-sourcing the GridVQA-X generation engine and paired diagnostic models, we provide a definitive, zero-ambiguity testbed to shift the field away from superficial plausibility metrics and toward the development of explainers capable of verifiably diagnosing relational cross-modal grounding.

Acknowledgements

We would like to thank the Precog Lab at IIIT Hyderabad for their valuable guidance and discussions throughout this work. In particular, we thank Vedanta S. P. and Debangan Mishra for their detailed feedback during multiple stages of the project. The views expressed are those of the authors and do not necessarily reflect the official policies or positions of the supporting organizations.

References

1. Agarwal, C.: Rethinking explainability in the era of multimodal ai (2025), https://arxiv.org/abs/2506.13060 4
2. Agarwal, C., Ley, D., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: Openxai: Towards a transparent evaluation of model explanations (2024), https://arxiv.org/abs/2206.11104 2, 4
3. Arras, L., Osman, A., Samek, W.: Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion **81**, 14–40 (May 2022). https://doi.org/10.1016/j.inffus.2021.11.008, http://dx.doi.org/10.1016/j.inffus.2021.11.008 2, 4
4. Chi, Z., Hou, Y., Pang, C., Cui, S., Akhtar, M., Sachan, M.: Chimera: Diagnosing shortcut learning in visual-language understanding (2025), https://arxiv.org/abs/2509.22437 3
5. Dang, Y., Huang, K., Huo, J., Yan, Y., Huang, S., Liu, D., Gao, M., Zhang, J., Qian, C., Wang, K., Liu, Y., Shao, J., Xiong, H., Hu, X.: Explainable and interpretable multimodal large language models: A comprehensive survey (2024), https://arxiv.org/abs/2412.02104 4, 19
6. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (Nov 2020). https://doi.org/10.1038/s42256-020-00257-z, http://dx.doi.org/10.1038/s42256-020-00257-z 19
7. Goldshmidt, R.: Attention, please! pixelshap reveals what vision-language models actually focus on (2025), https://arxiv.org/abs/2503.06670 2, 4, 20
8. Goldshmidt, R., Horovicz, M.: Tokenshap: Interpreting large language models with monte carlo shapley value estimation (2024), https://arxiv.org/abs/2407.10114 20
9. Hessel, J., Lee, L.: Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! (2020), https://arxiv.org/abs/2010.06572 2, 4, 12, 20
10. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering (2019), https://arxiv.org/abs/1902.09506 3
11. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning (2016), https://arxiv.org/abs/1612.06890 3
12. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr – modulated detection for end-to-end multi-modal understanding (2021), https://arxiv.org/abs/2104.12763 10, 19
13. Klein, L., Lüth, C.T., Schlegel, U., Bungert, T.J., El-Assady, M., Jäger, P.F.: Navigating the maze of explainable ai: A systematic approach to evaluating methods and metrics (2025), https://arxiv.org/abs/2409.16756 4
14. Le, P., Nauta, M., Nguyen, B., Pathak, S., Schlötterer, J., Seifert, C.: Benchmarking explainable ai - a survey on available toolkits and open challenges. pp. 6665–6673 (08 2023). https://doi.org/10.24963/ijcai.2023/747 2
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day (2023), https://arxiv.org/abs/2306.00890 2
16. Li, K., Vosselman, G., Yang, M.Y.: Multimodal rationales for explainable visual question answering (2025), https://arxiv.org/abs/2402.03896 3

17. Li, X., Du, M., Chen, J., Chai, Y., Lakkaraju, H., Xiong, H.: $\mathcal{M}^4$: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=6zcfrSz98y 2
18. Li, Z., Wen, X., Lou, J., Ji, Y., Lu, Y., Han, X., Zhang, D., Sun, L.: The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models (2025), https://arxiv.org/abs/2503.03122 19
19. Liang, P.P., Cheng, Y., Fan, X., Ling, C.K., Nie, S., Chen, R., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., Salakhutdinov, R., Morency, L.P.: Quantifying modeling multimodal interactions: An information decomposition framework (2023), https://arxiv.org/abs/2302.12247 2
20. Liang, P.P., Lyu, Y., Chhablani, G., Jain, N., Deng, Z., Wang, X., Morency, L.P., Salakhutdinov, R.: Multiviz: Towards visualizing and understanding multimodal models (2023), https://arxiv.org/abs/2207.00056 2, 4, 11, 20
21. Lyu, Y., Liang, P.P., Deng, Z., Salakhutdinov, R., Morency, L.P.: Dime: Fine-grained interpretations of multimodal models via disentangled local explanations (2022), https://arxiv.org/abs/2203.02013 2, 3, 4, 11, 20
22. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys **55**(13s), 1–42 (Jul 2023). https://doi.org/10.1145/3583558, http://dx.doi.org/10.1145/3583558 2
23. Rahmanzadehgervi, P., Bolton, L., Taesiri, M.R., Nguyen, A.T.: Vision language models are blind: Failing to translate detailed visual features into words (2025), https://arxiv.org/abs/2407.06581 19
24. Rodis, N., Sardianos, C., Radoglou-Grammatikis, P., Sarigiannidis, P., Varlamis, I., Papadopoulos, G.T.: Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions (2024), https://arxiv.org/abs/2306.05731 2
25. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations (2017), https://arxiv.org/abs/1703.03717 11, 19
26. Salewski, L., Koepke, A.S., Lensch, H.P.A., Akata, Z.: CLEVR-X: A Visual Reasoning Dataset for Natural Language Explanations, p. 69–88. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-04083-2_5, http://dx.doi.org/10.1007/978-3-031-04083-2_5 2
27. Shao, X., Skryagin, A., Stammer, W., Schramowski, P., Kersting, K.: Right for better reasons: Training differentiable models by constraining their influence functions. Proceedings of the AAAI Conference on Artificial Intelligence **35**(11), 9533–9540 (May 2021). https://doi.org/10.1609/aaai.v35i11.17148, https://ojs.aaai.org/index.php/AAAI/article/view/17148 19
28. Si, Q., Meng, F., Zheng, M., Lin, Z., Liu, Y., Fu, P., Cao, Y., Wang, W., Zhou, J.: Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 3698–3712. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.findings-emnlp.271, https://aclanthology.org/2022.findings-emnlp.271/ 19

29. Sun, S., An, W., Tian, F., Nan, F., Liu, Q., Liu, J., Shah, N., Chen, P.: A review of multimodal explainable artificial intelligence: Past, present and future (2024), https://arxiv.org/abs/2412.14056 4, 19
30. Tucci, R.R.: Introduction to judea pearl's do-calculus (2013), https://arxiv.org/abs/1305.5506 5
31. Wang, Z., Wang, K.: Multishap: A shapley-based framework for explaining cross-modal interactions in multimodal ai models (2026), https://arxiv.org/abs/2508.00576 2, 4, 11, 20
32. Wenderoth, L., Hemker, K., Simidjievski, N., Jamnik, M.: Measuring cross-modal interactions in multimodal models. Proceedings of the AAAI Conference on Artificial Intelligence **39**(20), 21501–21509 (Apr 2025). https://doi.org/10.1609/aaai.v39i20.35452, http://dx.doi.org/10.1609/aaai.v39i20.35452 2, 4, 12, 20
33. Williams, P.L., Beer, R.D.: Nonnegative decomposition of multivariate information (2010), https://arxiv.org/abs/1004.2515 21
34. Xu, Z., Xiang, X., Liang, Y.: Overcoming shortcut problem in vlm for robust out-of-distribution detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15402–15412 (2025). https://doi.org/10.1109/CVPR52734.2025.01435 19
35. Ye, W., Jiang, L., Xie, E., Zheng, G., Ma, Y., Cao, X., Guo, D., Qi, D., He, Z., Tian, Y., Coffee, M., Zeng, Z., Li, S., Ting-hao, Huang, Wang, Z., Rehg, J.M., Kautz, H., Zhang, A.: The clever hans mirage: A comprehensive survey on spurious correlations in machine learning (2025), https://arxiv.org/abs/2402.12715 19
36. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: 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), https://openreview.net/forum?id=KRLUvxh8uaX 3, 7
37. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2025), https://arxiv.org/abs/2303.00915 2
38. Zhang, Z., Yadav, S., Han, F., Shutova, E.: Cross-modal information flow in multimodal large language models (2025), https://arxiv.org/abs/2411.18620 3

A Detailed Related Work

A.1 Shortcut Learning in Vision-Language Models

Deep learning models frequently learn decision rules that exploit spurious correlations in training data rather than the intended causal logic. This phenomenon is often referred to as the *Clever Hans effect*, named after a horse that appeared to perform arithmetic but was later discovered to be responding to subtle cues from human observers rather than actually solving the task [35]. In machine learning, this effect describes models that appear to perform a task correctly while in reality relying on unintended signals or dataset artifacts.

In the context of Visual Question Answering (VQA), models often rely on language-only priors to predict answers without actually looking at the image [23]. Recent research identifies various shortcuts that artificially inflate performance on multimodal benchmarks: i) **Linguistic and Background Statistics:** Models frequently rely on background correlations rather than the primary object's features to make predictions [6, 34]; ii) **Keyword Shortcuts:** Models often learn a shallow, direct mapping between a noun, color, or shape and a high-probability answer, entirely bypassing the relational or spatial context [18, 28].

While these shortcuts yield high accuracy on identically distributed test sets, they cause catastrophic failure in out-of-distribution or adversarial scenarios. Our work directly addresses this by mathematically isolating these keyword and background shortcuts within the GridVQA-X generation algorithm.

A.2 Explainability Guided Training

Existing research [25, 27] highlights that achieving high accuracy is insufficient if the underlying decision-making process relies on spurious correlations. Explainability guided training addresses this by incorporating auxiliary constraints during optimization to align model reasoning with human expectations. In vision-language domains, this often takes the form of visual grounding supervision, where models are penalized if their attention or predicted bounding boxes do not align with ground-truth causal features [12].

By explicitly supervising the intermediate representations, researchers can mitigate shortcut reliance. In the context of GridVQA-X, we utilize this paradigm not just to improve accuracy, but as a controlled mechanism to train our diagnostic models. By modulating the grounding loss and dynamically weighting the cross-entropy loss, we successfully enforce the behavioral divergence required to train both M_{pure} (which strictly aligns with causal spatial-relational geometry) and M_{spur} (which bypasses spatial grounding to exploit injected statistical traps).

A.3 Detailed Mechanics of Evaluated MxAI Methods

Existing research [5, 29] highlights the growing need for explainability techniques. We broadly classify these methods based on their attribution scope, dividing them into local (micro-level) and global (macro-level) taxonomies:

1. Local (Micro-Level) Attribution Methods. These methods focus on fine-grained attributions, seeking to explain model predictions by isolating the specific contributions of individual image patches, pixels, or text tokens, as well as the localized interactions between them.

- **MultiSHAP [31]:** A model-agnostic framework that leverages the Shapley Interaction Index to attribute predictions to pairwise interactions between fine-grained visual and textual elements (such as image patches and text tokens).
- **Dime [21]:** DIME enables fine-grained analysis by explicitly disentangling model behavior into unimodal contributions (UC) and multimodal interactions (MI). It is designed to maintain generality across arbitrary modalities, model architectures, and tasks, allowing stakeholders to understand model behavior and performing debugging.
- **MultiViz [20]:** A framework for visualizing and analyzing model behavior by scaffolding interpretability into four stages: (1) unimodal importance, (2) cross-modal interactions, (3) multimodal representations, and (4) multimodal prediction. These complementary stages enable users to simulate predictions, assign interpretable concepts to features, perform error analysis, and debug models.
- **PixelSHAP + TokenSHAP [7, 8]:** A model-agnostic multimodal attribution framework that combines TokenSHAP for textual analysis and PixelSHAP for visual reasoning to produce joint explanations for vision–language model predictions. TokenSHAP attributes the model's output to individual tokens or substrings within the input prompt using Shapley values, modeling each token as a cooperative player whose contribution is estimated through Monte Carlo sampling over token subsets. PixelSHAP extends the same Shapley-value framework to visual inputs by treating segmented image regions or objects as players in the cooperative game and quantifying their influence on the model output through systematic perturbations of these visual components.

2. Global (Macro-Level) Attribution Methods. Rather than focusing on individual tokens or patches, these methods evaluate the overall contribution of entire modalities and the high-level synergistic interactions between the visual and textual inputs as a whole.

- **InterSHAP [32]:** A game-theoretic approach that shifts the focus from local tokens to the overall contribution of entire modalities, specifically isolating the synergistic interactions between the visual and textual inputs as a whole.
- **EMAP [9]:** A diagnostic tool designed to identify whether cross-modal interactions genuinely improve model performance. It utilizes an empirical multimodal additive function projection to modify model predictions such that cross-modal interactions are eliminated, isolating the additive, unimodal structure. This reveals whether a high-performing black-box model actually utilizes complex cross-modal reasoning or mostly exploits unimodal signals in the data.

– **Partial Information Decomposition (PID) [33]:** An information-theoretic framework for analyzing how multiple input sources jointly contribute to a model's prediction. PID decomposes the mutual information between a set of inputs and an output into distinct components representing unique, redundant, and synergistic information. In multimodal settings, this decomposition enables quantifying how much predictive information is provided exclusively by the visual modality, exclusively by the textual modality, redundantly by both, or only through their joint interaction. PID provides a principled way to measure whether a multimodal model truly relies on cross-modal reasoning rather than unimodal shortcuts.

B Dataset Generation Details and Proofs

B.1 Case-0: Answer Prior Bias Elimination

As introduced in Section 3.3, the Answer Prior Bias (Case-0) heuristic assumes the model ignores the visual modality entirely and predicts $\hat{y}$ by exploiting the marginal distribution of answers $P(\mathcal{Y})$.

In natural VQA, answers are heavily skewed, allowing models to achieve artificially high accuracy via prior guessing. In GridVQA-X, we mitigate this through explicit dataset balancing. As verified by our dataset statistics, for *Form 1* (Existence) queries, the generative engine enforces strict target-distractor independence, yielding a near-perfect balanced distribution across all buckets.

For *Form 0* (Counting) queries, placing multiple objects satisfying complex spatial constraints on a finite grid inherently introduces a structural skew. For instance, high depth queries in sparse environments ($D3_M_F0_d_{0.3}$) yield "1" as the most frequent answer (37.2%) (Fig. 5). Thus, a pure Case-0 heuristic is strictly bounded by this empirical ceiling, making it insufficient to solve the dataset.

Our generation algorithm guarantees that the marginal answer distributions are almost identical across the two splits: $P(\mathcal{Y})_{Pure} = P(\mathcal{Y})_{Spurious}$. Since the Case-0 heuristic space is almost identical between the two environments, prior bias cannot account for any downstream divergence in behavior between the Pure and Spurious testbeds.

B.2 Case-1: The Spatial Shortcut (Bag-of-Words) Elimination

Our empirical validation confirms the mathematical design of the dataset splits regarding the spatial shortcut. Predicting the target count based solely on the bag-of-words spatial shortcut yields a 100% success rate ($P = 1.00$) in $\mathcal{D}_{\text{spur}}$, successfully embedding the behavioral trap. Conversely, in $\mathcal{D}_{\text{pure}}$, the deterministic injection of adversarial distractors collapses this shortcut's success rate to 36.07%, proving that a model mathematically cannot converge on this split without computing cross-modal spatial synergy.

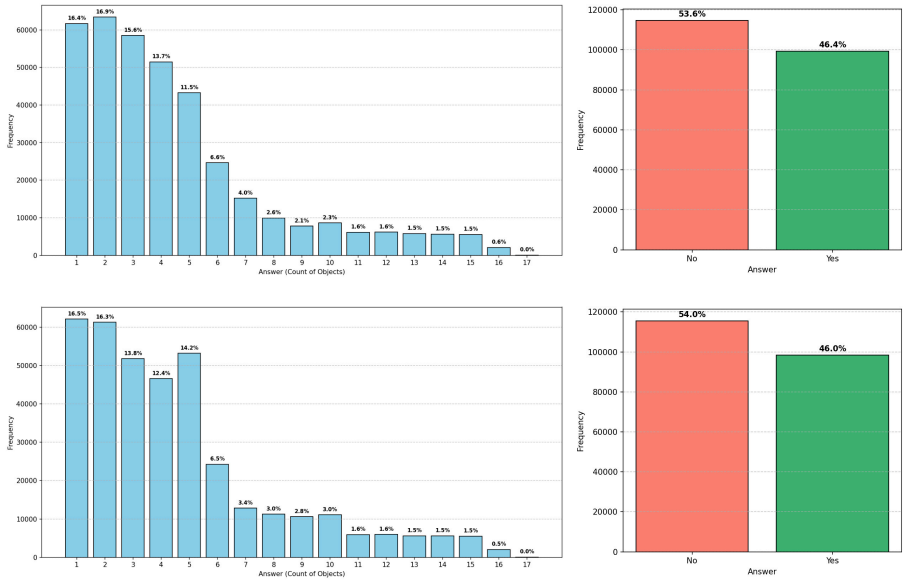

Fig. 5: Answer distribution in the training set. Top row shows the answer distribution for the $\mathcal{D}_{\text{pure}}$, while the bottom row shows the answer distribution for the $\mathcal{D}_{\text{spur}}$. Left plots correspond to the counting task (Form0) with answers ranging from 1–17, and right plots correspond to the binary yes/no task (Form1).

B.3 Case-2: Visual Feature Dominance Elimination

The Visual Feature Dominance heuristic assumes the model ignores the textual query $\mathcal{Q}$ entirely and predicts the answer based solely on layout imbalances inherent in the image $\mathcal{V}$. Two primary sub-hypotheses exist within this family: the Salience Prior (H_{sal}), where the presence of a specific attribute correlates with an answer, and the Majority Class Prior (H_{maj}), where the model blindly predicts based on the most frequent object type.

To prevent any single attribute from becoming a predictive shortcut (H_{sal}), our generation pipeline utilizes independent, uniform multi-randomization. During semantic instantiation, the exact values for color, shape, and direction are drawn from uniform distributions. Because the marginal probability of any specific visual feature occurring is statistically decorrelated from the ground-truth answer, H_{sal} provides no reliable predictive signal.

To nullify the visual majority dominance (H_{maj}), the generative engine must logically decorrelate the target from the visual majority, satisfying Property 2 (Orthogonality of Frequency). We structurally counter the majority heuristic through two distinct mechanisms.

First, for all buckets in $\mathcal{D}_{\text{spur}}$ and the non-relational Attribute-Only buckets in $\mathcal{D}_{\text{pure}}$, the generator intervenes during noise injection. As outlined in Algo-

rithm 2, it utilizes a stochastic branching mechanism parameterized over the atomic feature space.

Algorithm 2 Stochastic Target-Distractor Balancing

Require: Target attribute set $\mathcal{A}_{tgt}$, Total grid capacity C
$u \sim \mathcal{U}(0,1)$ ▷ Sample from standard uniform distribution
if $u < 0.5$ **then**
▷ **BOOST Branch:** Force target to be the majority class
    $N_{tgt} \leftarrow$ Sample(Upper Range)
    Inject($N_{tgt}, \mathcal{A}_{tgt}$)
else
▷ **SUPPRESS Branch:** Force a disjoint distractor to be the majority
    $N_{tgt} \leftarrow$ Sample(Lower Range)
    $\mathcal{A}_{dist} \leftarrow$ Sample($\mathcal{V}_{attributes} \setminus \mathcal{A}_{tgt}$)
    $N_{dist} \leftarrow$ Sample($> N_{tgt}$)
    Inject($N_{tgt}, \mathcal{A}_{tgt}$)
    Inject($N_{dist}, \mathcal{A}_{dist}$)
end if

While the theoretical split is uniform, downstream generative constraints (*e.g.* valid grid space exhaustion and integer bounding) modulate this distribution. This active balancing bounds the success rate of the H_{maj} heuristic to an empirical average of $\sim 44\%$ across evaluated buckets, rendering it statistically insufficient.

Second, for the relational spatial queries in $\mathcal{D}_{\text{pure}}$, the H_{maj} heuristic is destroyed as an emergent consequence of eliminating the spatial shortcut (Case-1). By injecting adversarial distractors that share the target's visual attributes but explicitly lie outside the valid bounding region, the target's visual class almost always becomes the majority class (*e.g.* 99.36% frequency in $D1_CO_F0_d_{0.3}$). However, because these distractors violate the spatial constraint, the global count of the majority class inherently exceeds the true relational ground-truth count. A model attempting to use H_{maj} will consistently overcount and fail, safely eliminating the heuristic.

B.4 Case-3: Logical Decomposition (Partial Logic) Elimination

The Logical Decomposition heuristic occurs when a model receives a multi-hop compositional query (e.g., Depth 2 or Depth 3) but evaluates only a subset of the spatial constraints. Let a complex query require the logical intersection of M spatial constraints relative to M anchors: $\mathcal{Q} = \bigwedge_{i=1}^{M} C_i$. A model exploiting partial logic will drop constraint k, effectively treating the intersection as $\bigwedge_{i \neq k} C_i$.

In $\mathcal{D}_{\text{spur}}$, the model never encounters this heuristic because it collapses entirely to the Case-1 Bag-of-Words shortcut, bypassing spatial evaluation altogether. However, in $\mathcal{D}_{\text{pure}}$, where Case-1 is mathematically eliminated, the model is forced to evaluate spatial regions. If the dataset allows the set of objects satisfying $\bigwedge_{i \neq k} C_i$ to frequently be a subset of those satisfying C_k, the model will learn to safely ignore anchor k without penalty.

To definitively eliminate this heuristic in the Pure dataset, we generalize the Robust Intersection requirement to M dimensions (Theorem 1). The formal proof expanding on the main text's proof sketch is as follows.

Proof of Theorem 1. Let Y_{true} be the set of valid target objects that correctly answer the multimodal query. By definition, these objects must possess the target visual attributes and reside entirely within the strictly valid spatial intersection of all anchors:

$$V_{\text{target}} = \bigcap_{i=1}^{M} V_i$$

A model employing the partial logic heuristic drops spatial constraint k to simplify computation, thus evaluating the relaxed geometric region:

$$V_{-k} = \bigcap_{i \neq k} V_i$$

Let Y_{partial} be the candidate set of objects detected by this heuristic. Because the intersection of a subset of regions is always a superset of the intersection of all regions, it geometrically holds that $V_{\text{target}} \subseteq V_{-k}$. Consequently, $Y_{\text{true}} \subseteq Y_{\text{partial}}$.

The set of false positives detected by the heuristic is strictly defined by the relative complement of these spatial regions:

$$Y_{\text{FP}} \subset (V_{-k} \setminus V_{\text{target}})$$

We can simplify the false positive bounding region as follows:

$$V_{-k} \setminus V_{\text{target}} = V_{-k} \setminus (V_{-k} \cap V_k) = V_{-k} \setminus V_k = C_k$$

Thus, any false positive objects must reside within the confuser region C_k. By algorithmically guaranteeing that C_k is non-empty and contains at least one adversarial object d_{adv} matching the target's visual attributes, the heuristic will blindly detect it, guaranteeing that $|Y_{\text{FP}}| \geq 1$.

Therefore, the total count of objects detected by the heuristic is:

$$|Y_{\text{partial}}| = |Y_{\text{true}}| + |Y_{\text{FP}}| > |Y_{\text{true}}|$$

Because the heuristic strictly overcounts the true targets, it yields an incorrect prediction, incurs a strict training loss penalty, and is mathematically eliminated. ■

Adversarial Confuser Placement Implementation. Rather than relying on random rejection sampling to fulfill this theorem, the $\mathcal{D}_{\text{pure}}$ generation engine actively constructs these adversarial confuser regions during the target placement phase. As detailed in Algorithm 3, the engine computes C_k for every anchor in the query.

Algorithm 3 Adversarial Confuser Placement for Depth ≥ 2

Require: List of Anchor Valid Regions $[V_1, V_2, \ldots, V_M]$, Target Attributes $\mathcal{A}_{tgt}$

▷ **Phase 1: Guaranteed Confuser Placement**
for $k = 1 \ldots M$ **do**
  $V_{-k} \leftarrow \bigcap_{i \neq k} V_i$ ▷ Intersection of all OTHER anchors
  $C_k \leftarrow V_{-k} \setminus V_k$ ▷ Subtract the dropped anchor's region
  if $C_k \neq \emptyset$ **then**
    $pos \leftarrow \text{Sample}(C_k)$
    $\text{Inject}(pos, \mathcal{A}_{tgt})$ ▷ Place target-attribute object in confuser region
  end if
end for
▷ **Phase 2: Stochastic Filler Overload**
while Noise Target Not Met **do**
  if random() < 0.75 **and** Sample From Outside Valid Region **then**
    $k \leftarrow \text{SampleRandomAnchor}()$
    $\text{Inject}(\text{Sample}(C_k), \mathcal{A}_{tgt})$
  end if
end while

Before injecting general background noise, the engine guarantees the placement of exactly one target-attribute object into every valid confuser region C_k. During the subsequent filler loop, the engine overwhelmingly (75% probability) targets these specific C_k regions when injecting adversarial spatial distractors.

Our empirical reports rigorously validate this active confuser placement. When a heuristic drops a spatial constraint, its predictive success rate plummets. For instance, in dense Depth 2 environments ($D2_M_F0_d_{0.7}$), dropping either anchor 1 or anchor 2 yields a predictive ratio of only 16.40% and 16.87%, respectively. Consequently, minimizing the training loss requires the strict evaluation of the full compositional intersection $\bigwedge_{i=1}^{M} C_i$, effectively neutralizing the Logical Decomposition heuristic.

B.5 Collision-Free Object Placement Requirements

To ensure that spatial relationships are unambiguous and C_k regions are geometrically sound, the generator employs a strictly collision-free placement algorithm. When instantiating the set of anchor objects $\mathcal{A}$ and target objects $\mathcal{T}_{\text{tgt}}$, the grid is modeled as a discrete 2D matrix. Anchors are placed sequentially; for multi-hop queries ($D > 1$), subsequent anchors are placed such that their valid intersecting region $V_{\text{target}} = \bigcap V_i$ contains at least one available grid cell. To prevent edge-case spatial ambiguities, we enforce a strict minimum grid-cell distance margin between objects involved in directional relationships.

C Dataset Details

Figures 6, 7, 8, and 9 illustrate representative examples from the $\mathcal{D}_{\text{pure}}$ and $\mathcal{D}_{\text{spur}}$ datasets across different reasoning depths, forms, question types, and scene

densities. Each panel shows an image–question pair along with the corresponding answer.

D Additional Results

D.1 Model Accuracies

Refer to Tables 6 and 7 for the accuracies of existing open-source models compared to our trained models, $\mathcal{M}_{\text{pure}}$ and $\mathcal{M}_{\text{spur}}$.

D.2 Local Explainability Algorithms

Tables 8, 9, 10, and 11 give the bucket-wise performance of the three algorithms on the dataset.

D.3 Global Explainability Algorithms

Figures 10, 11, 12, and 13 illustrate attribution mass distributions and interaction-based explanations for both $\mathcal{M}_{\text{pure}}$ and $\mathcal{M}_{\text{spur}}$ across datasets.

E Limitations and Future Work

While GridVQA-X mathematically guarantees unique ground-truth explanations, it operates within a highly controlled 2D abstraction that fundamentally lacks the noisy feature distributions and deeply entangled semantic representations of real-world multimodal tasks. Additionally, the framework's scope is currently restricted to spatial-relational composition and basic attribute binding. It does not measure an explainer's faithfulness when dealing with other complex cross-modal synergies, such as temporal reasoning, physical commonsense, or complex mathematical logic.

To bridge this abstraction gap, future work should extend these diagnostic principles to richer spatial reasoning benchmarks to provide a more rigorous stress test involving continuous spatial relations and realistic object interactions. Furthermore, since our empirical results expose that existing explainers often devolve into shallow detectors or hallucinate complex interactions, future research must shift toward developing novel explanation algorithms specifically designed to pass this zero-ambiguity testbed. Finally, extending the current evaluation pipeline to explicitly assess the faithfulness of natural language rationales and chain-of-thought explanations generated directly by multimodal LLMs presents a highly promising direction.

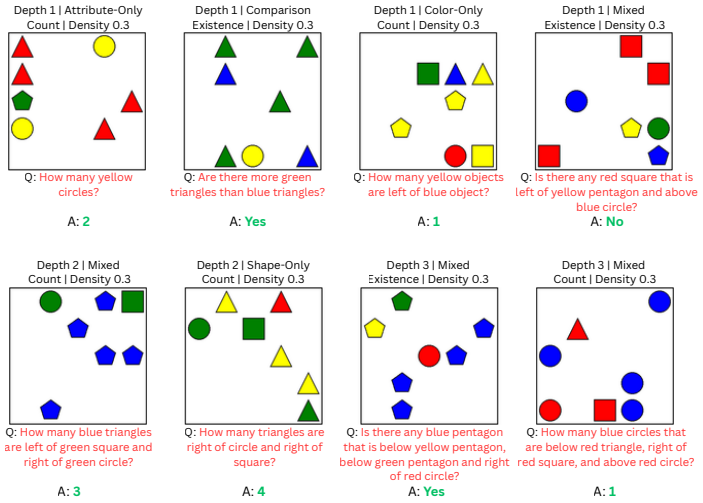

Fig. 6: Example samples from $\mathcal{D}_{\text{pure}}$ with density $d = 0.3$.

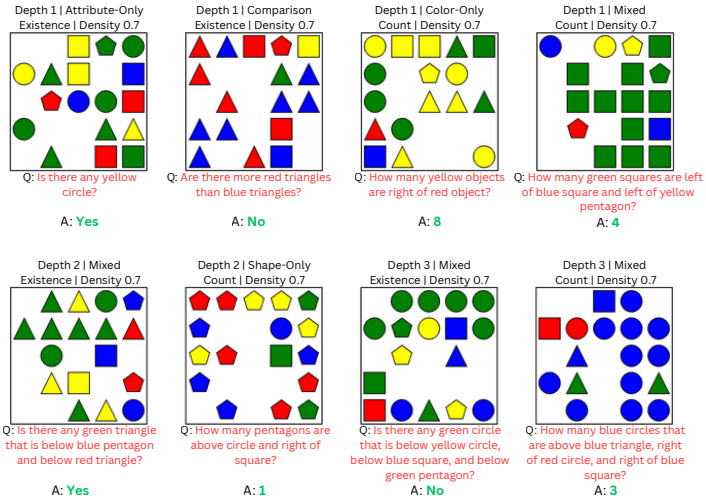

Fig. 7: Example samples from $\mathcal{D}_{\text{pure}}$ with density $d = 0.7$.

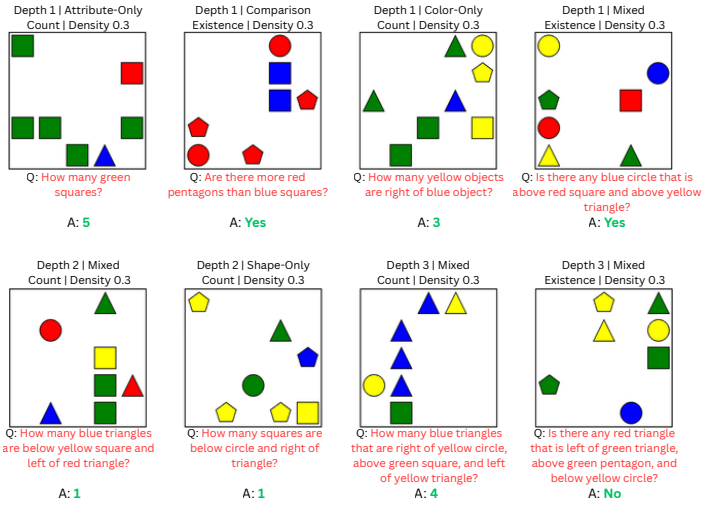

Fig. 8: Example samples from $\mathcal{D}_{\text{spur}}$ with density $d = 0.3$. Note that the correct answer can be inferred even when the directional cues are ignored.

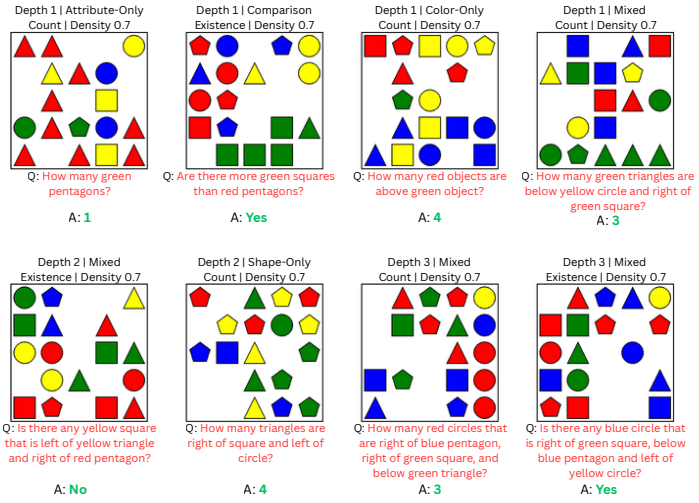

Fig. 9: Example samples from $\mathcal{D}_{\text{spur}}$ with density $d = 0.7$. Note that the correct answer can be inferred even when the directional cues are ignored.

Table 6: Model Accuracies for Open Source Models on $\mathcal{D}_{\text{pure}}$ with $d_{0.3}$

Bucket	Depth	Qwen2.5-VL-3B	Gemma-3-4B	Llava-1.5-7B	Qwen3-VL-30B	$\mathcal{M}_{\text{spur}}$	$\mathcal{M}_{\text{pure}}$
Form 0							
A	D1	68.0	48.0	16.0	70.0	100	**100**
SO	D1	20.0	18.0	18.0	24.0	32	**100**
SO	D2	22.0	34.0	22.0	32.0	16	**100**
CO	D1	34.0	20.0	14.0	28.0	18	**100**
CO	D2	22.0	38.0	22.0	38.0	16	**100**
M	D1	32.0	24.0	22.0	20.0	36	**100**
M	D2	28.0	30.0	28.0	28.0	8	**100**
M	D3	32.0	38.0	40.0	34.0	14	**100**
CMP	D1	76.0	50.0	52.0	72.0	100	**100**
CMP	D2	66.0	44.0	54.0	66.0	96	**100**
CMP	D3	46.0	42.0	38.0	50.0	96	**100**
Form 1							
A	D1	94.0	90.0	50.0	94.0	100	**100**
SO	D1	48.0	44.0	60.0	80.0	46	**100**
SO	D2	58.0	64.0	56.0	60.0	60	**100**
CO	D1	56.0	52.0	58.0	78.0	50	**100**
CO	D2	62.0	60.0	58.0	74.0	56	**100**
M	D1	64.0	62.0	54.0	70.0	62	**100**
M	D2	48.0	48.0	46.0	68.0	46	**100**
M	D3	56.0	56.0	58.0	40.0	56	**100**

Table 7: Model Accuracies for Open Source Models on $\mathcal{D}_{\text{pure}}$ with $d_{0.7}$

Bucket	Depth	Qwen2.5-VL-3B	Gemma-3-4B	Llava-1.5-7B	Qwen3-VL-30B	$\mathcal{M}_{\text{spur}}$	$\mathcal{M}_{\text{pure}}$
Form 0							
A	D1	30.0	14.0	14.0	36.0	100	**100**
SO	D1	22.0	10.0	10.0	12.0	2	**100**
SO	D2	18.0	22.0	18.0	18.0	0	**100**
CO	D1	22.0	4.0	0.0	6.0	6	**100**
CO	D2	12.0	20.0	22.0	10.0	2	**100**
M	D1	16.0	8.0	6.0	14.0	6	**100**
M	D2	12.0	22.0	28.0	10.0	2	**100**
M	D3	32.0	24.0	30.0	10.0	0	**100**
CMP	D1	66.0	42.0	42.0	76.0	**100**	98
CMP	D2	52.0	14.0	48.0	36.0	92	**100**
CMP	D3	28.0	18.0	24.0	20.0	80	**100**
Form 1							
A	D1	92.0	70.0	48.0	88.0	100	**100**
SO	D1	46.0	46.0	60.0	68.0	46	**100**
SO	D2	62.0	60.0	56.0	48.0	60	**100**
CO	D1	50.0	50.0	54.0	70.0	50	**100**
CO	D2	56.0	56.0	56.0	56.0	56	**100**
M	D1	58.0	56.0	60.0	78.0	56	**100**
M	D2	50.0	52.0	48.0	66.0	46	**100**
M	D3	58.0	44.0	54.0	50.0	56	**100**

Table 8: RMA (↑) comparison on M_{pure} with $\mathcal{D}_{\text{pure}}$. Note the severe performance drop across all methods on Existence tasks (Form 1) compared to Counting tasks (Form 0), highlighting a structural reliance on mask volume rather than precise logical routing.

		DIME		MultiSHAP		MultiViz	
Bucket	Depth	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$
Form 0							
A	D1	26.7± 9.8	41.4± 13.6	66.7± 29.9	69.2± 21.6	47.3± 15.1	56.4± 17.0
SO	D1	25.8± 10.1	36.1± 15.6	73.5± 20.5	61.4± 23.2	51.3± 11.7	48.6± 14.4
SO	D2	26.3± 8.2	30.5± 12.9	64.7± 9.9	49.5± 17.9	49.7± 11.8	45.4± 13.6
CO	D1	26.9± 9.6	40.4± 14.3	70.2± 14.4	66.7± 12.7	47.9± 13.3	50.7± 14.1
CO	D2	27.6± 7.3	32.1± 12.6	68.1± 15.7	54.2± 15.7	48.0± 11.8	51.8± 12.4
M	D1	29.6± 9.7	35.9± 14.4	60.4± 24.0	61.4± 17.7	47.9± 14.6	45.6± 11.4
M	D2	31.2± 9.5	30.0± 11.4	70.4± 10.8	53.8± 14.5	47.6± 12.7	46.7± 14.8
M	D3	31.1± 9.4	27.9± 7.0	68.3± 10.9	47.1± 15.4	50.8± 13.5	44.9± 9.8
CMP	D1	35.8± 4.9	56.0± 12.5	87.4± 5.8	79.2± 8.4	46.0± 13.6	55.7± 13.2
CMP	D2	34.2± 5.9	44.0± 10.4	83.9± 4.0	71.6± 12.5	49.6± 11.7	49.6± 15.7
CMP	D3	35.4± 4.8	38.2± 7.6	82.6± 3.3	62.0± 9.0	52.2± 10.2	52.4± 11.4
Form 1							
A	D1	23.1± 9.7	30.4± 15.2	56.4± 15.9	52.9± 22.2	42.9± 18.5	48.3± 25.1
SO	D1	12.6± 8.6	16.5± 17.0	45.1± 20.9	32.4± 16.3	28.6± 16.5	31.3± 19.4
SO	D2	17.1± 6.7	18.2± 9.1	50.3± 15.5	35.9± 11.9	37.8± 13.6	30.5± 13.4
CO	D1	16.3± 10.3	19.0± 13.5	44.9± 19.2	28.6± 10.3	32.1± 13.5	30.6± 18.4
CO	D2	20.8± 9.9	20.1± 9.7	51.4± 16.9	29.7± 9.1	40.4± 14.6	36.8± 11.1
M	D1	17.4± 9.4	19.1± 13.2	43.3± 20.5	31.9± 16.5	34.8± 18.4	31.7± 16.4
M	D2	20.8± 7.4	18.7± 6.8	37.3± 13.4	34.2± 12.8	36.1± 13.8	39.1± 19.4
M	D3	24.3± 6.8	24.3± 8.1	43.7± 16.2	36.4± 9.4	44.9± 11.7	40.8± 13.5

Table 9: IoU (↑) comparison on M_{pure} with $\mathcal{D}_{\text{pure}}$. MultiSHAP demonstrates strong object grounding, while DIME exhibits poor precision due to systemic attribution spread, and MultiViz suffers a complete spatial grounding collapse (< 10% across all depths).

		DIME		MultiSHAP		MultiViz	
Bucket	Depth	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$
Form 0							
A	D1	28.6± 14.4	27.4± 10.5	71.3± 31.6	62.1± 22.8	7.7± 3.7	5.3± 3.4
SO	D1	20.7± 10.1	25.2± 10.2	38.7± 25.8	44.3± 25.7	4.6± 3.2	3.9± 3.1
SO	D2	21.4± 8.0	22.2± 7.9	55.7± 10.7	35.9± 21.9	4.0± 2.7	4.1± 2.9
CO	D1	21.0± 9.2	23.7± 9.8	68.7± 12.3	58.4± 20.2	3.1± 2.4	2.2± 2.0
CO	D2	19.6± 6.4	21.8± 8.2	53.7± 18.7	56.8± 11.5	3.6± 2.3	2.9± 2.5
M	D1	25.9± 10.1	23.1± 9.6	51.1± 26.1	64.0± 13.1	5.2± 2.8	3.5± 3.1
M	D2	22.7± 8.3	20.5± 7.3	59.1± 21.4	60.1± 15.3	4.4± 2.9	3.6± 2.2
M	D3	20.1± 6.7	18.7± 7.8	47.7± 17.6	45.6± 19.6	4.6± 2.6	4.3± 2.3
CMP	D1	28.8± 7.0	27.4± 8.5	84.0± 19.5	47.0± 20.3	4.8± 2.7	3.5± 1.8
CMP	D2	24.4± 5.8	25.5± 7.3	66.2± 21.5	51.7± 20.2	4.0± 2.3	2.4± 1.6
CMP	D3	20.3± 5.3	20.8± 7.4	48.6± 14.0	49.5± 11.8	4.3± 2.1	2.9± 2.1
Form 1							
A	D1	31.3± 14.4	28.0± 13.1	73.0± 9.4	58.7± 30.2	9.0± 3.4	7.3± 4.8
SO	D1	11.7± 8.2	12.7± 9.4	42.3± 16.3	41.3± 20.6	5.0± 3.8	6.3± 4.1
SO	D2	14.5± 6.9	15.6± 6.6	37.0± 12.0	31.2± 15.6	4.8± 2.8	5.6± 3.1
CO	D1	15.7± 10.0	18.8± 14.7	47.4± 16.9	41.5± 27.6	5.6± 4.3	6.0± 4.0
CO	D2	14.4± 7.1	14.6± 7.6	47.4± 19.4	39.6± 15.3	4.4± 3.3	5.0± 3.2
M	D1	17.7± 11.5	20.1± 14.1	44.0± 21.5	33.4± 20.6	6.5± 4.7	6.9± 4.6
M	D2	15.5± 6.6	15.6± 7.8	33.6± 12.2	43.0± 19.9	5.6± 2.4	5.2± 3.0
M	D3	17.3± 6.3	17.4± 6.7	47.4± 18.6	39.0± 10.2	4.2± 2.5	4.3± 2.7

Table 10: RMA (↑) comparison on $M_{\mathbf{spur}}$ with $\mathcal{D}_{\mathbf{spur}}$. MultiSHAP scores noticeably higher on this shortcut-reliant model than on the faithful M_{pure}, showing that game-theoretic methods may inherently favor shallow 1-to-1 feature matching over entangled spatial synergy.

		DIME		MultiSHAP		MultiViz	
Bucket	Depth	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$
Form 0							
A	D1	23.5± 7.0	41.2± 15.0	75.2± 20.7	85.2± 4.7	36.9± 20.5	44.9± 18.1
SO	D1	24.0± 9.0	37.9± 14.5	83.3± 5.1	78.7± 17.4	37.6± 17.0	36.1± 16.4
SO	D2	26.0± 8.0	34.7± 14.3	77.7± 10.8	69.3± 17.0	49.4± 18.5	53.6± 18.6
CO	D1	26.6± 7.5	36.2± 13.9	77.7± 12.8	80.7± 7.8	47.5± 17.2	45.1± 12.0
CO	D2	25.4± 7.8	33.1± 10.4	78.3± 7.9	81.0± 10.2	53.0± 13.8	51.8± 12.6
M	D1	26.3± 7.5	37.5± 11.9	71.4± 18.3	64.8± 17.5	47.8± 15.4	49.2± 15.9
M	D2	27.6± 8.0	33.0± 9.8	71.4± 13.0	64.8± 16.5	57.5± 15.6	50.8± 13.3
M	D3	32.5± 6.1	34.5± 7.9	72.9± 13.3	57.2± 19.2	58.5± 15.4	43.7± 12.5
CMP	D1	32.4± 8.0	51.6± 13.9	82.0± 7.8	71.0± 12.5	47.3± 17.6	49.1± 19.4
CMP	D2	31.6± 5.6	44.4± 13.0	79.8± 6.7	63.1± 15.0	53.4± 14.5	47.4± 14.9
CMP	D3	33.6± 7.6	38.2± 9.7	79.9± 4.1	62.5± 11.3	54.7± 12.5	46.3± 14.4
Form 1							
A	D1	19.4± 9.1	32.4± 17.5	52.1± 16.8	59.5± 26.2	24.7± 15.5	28.6± 17.4
SO	D1	10.3± 8.6	17.1± 21.1	56.9± 22.2	48.5± 21.9	33.6± 21.5	30.7± 21.6
SO	D2	14.2± 7.8	14.7± 7.8	71.3± 14.3	61.8± 19.9	51.8± 21.9	45.4± 16.7
CO	D1	14.7± 9.2	17.5± 13.9	56.4± 22.0	43.7± 22.5	23.2± 22.7	21.2± 19.6
CO	D2	18.6± 9.6	21.1± 12.8	62.0± 12.7	56.5± 11.5	44.2± 20.5	34.1± 16.7
M	D1	16.6± 10.2	19.8± 15.8	44.5± 25.4	26.5± 17.7	26.3± 17.4	22.7± 15.7
M	D2	17.3± 8.4	16.8± 10.8	46.3± 12.8	29.4± 18.0	39.9± 17.1	25.2± 14.7
M	D3	23.8± 8.0	25.9± 13.7	59.3± 21.1	36.9± 14.1	43.7± 18.1	30.7± 13.6

Table 11: IoU (↑) comparison on M_{spur} with $\mathcal{D}_{\text{spur}}$. MultiSHAP isolates the independent visual features driving the Bag-of-Words shortcut, whereas DIME and MultiViz consistently fail to generate coherent, object-centric bounding boxes regardless of the reasoning pathway.

		DIME		MultiSHAP		MultiViz	
Bucket	Depth	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$	$d_{0.3}$	$d_{0.7}$
Form 0							
A	D1	25.1± 12.6	25.1± 9.1	66.1± 20.6	56.7± 24.5	9.3± 5.1	5.1± 3.9
SO	D1	20.2± 8.8	22.9± 9.1	44.3± 23.1	58.8± 19.7	7.6± 4.4	4.3± 3.8
SO	D2	20.5± 8.1	23.3± 10.1	48.0± 20.4	33.4± 17.2	8.2± 4.3	6.7± 3.9
CO	D1	21.6± 8.1	24.0± 7.8	61.5± 16.3	64.8± 14.2	6.3± 4.0	6.4± 4.1
CO	D2	16.6± 5.8	19.7± 6.7	46.7± 21.6	63.7± 11.9	7.1± 3.3	6.8± 3.3
M	D1	21.4± 8.8	23.5± 8.4	50.9± 18.4	47.9± 17.8	7.4± 4.3	6.0± 4.0
M	D2	18.6± 7.3	20.9± 8.1	40.9± 21.1	36.0± 21.9	7.9± 3.8	5.8± 3.2
M	D3	16.1± 6.1	18.0± 6.5	44.7± 21.7	35.4± 17.4	6.3± 3.9	4.0± 2.8
CMP	D1	26.1± 9.2	28.0± 9.1	54.7± 21.8	40.3± 18.2	6.5± 3.0	3.8± 3.5
CMP	D2	21.4± 6.9	22.8± 10.3	61.4± 19.0	36.6± 17.1	6.5± 3.8	4.4± 2.5
CMP	D3	15.3± 5.2	18.4± 7.3	44.9± 17.2	34.2± 14.7	6.8± 3.3	4.5± 2.7
Form 1							
A	D1	24.6± 15.5	26.7± 14.0	69.0± 19.6	39.0± 20.6	7.4± 6.2	8.2± 6.3
SO	D1	9.4± 8.5	12.1± 12.1	62.3± 31.9	42.6± 24.0	9.3± 6.6	9.8± 5.6
SO	D2	11.5± 6.9	12.4± 6.6	45.2± 23.6	48.6± 17.7	8.6± 5.1	10.1± 4.4
CO	D1	13.6± 9.6	15.0± 12.2	61.6± 23.9	42.8± 19.5	5.0± 5.0	5.4± 5.3
CO	D2	12.2± 5.9	12.4± 8.7	57.2± 16.7	48.8± 15.7	6.5± 3.8	7.4± 4.5
M	D1	15.0± 10.0	14.3± 11.1	34.7± 22.6	20.6± 19.5	6.6± 5.0	6.9± 4.8
M	D2	12.1± 6.9	10.3± 5.7	37.5± 17.6	20.9± 10.9	7.0± 4.0	6.2± 3.6
M	D3	14.2± 6.2	13.5± 7.0	33.3± 20.4	27.6± 15.4	6.7± 3.5	6.3± 3.6

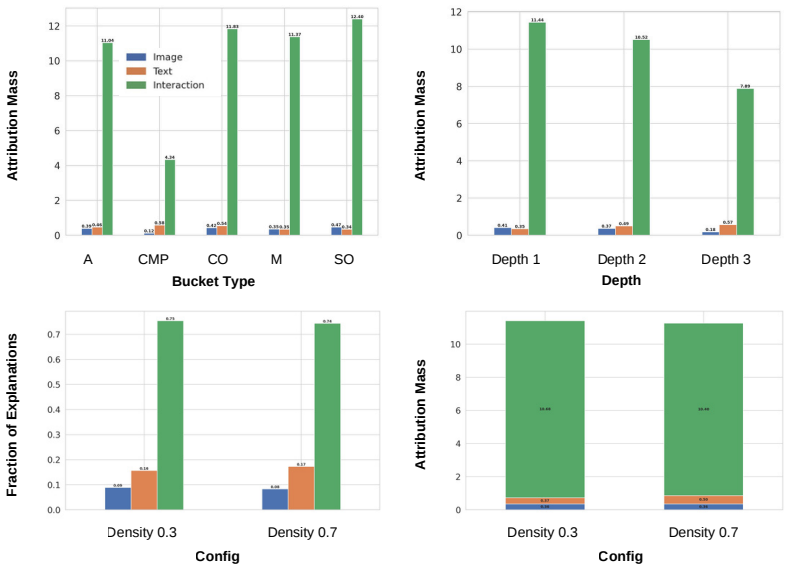

Fig. 10: EMAP attribution mass and explanation fractions for $\mathcal{M}_{\text{pure}}$ evaluated on $\mathcal{D}_{\text{pure}}$, segmented by query type, relational depth, and grid density.

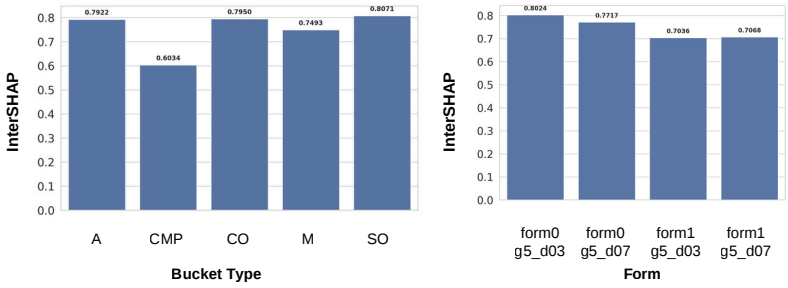

Fig. 11: InterSHAP scores on $\mathcal{M}_{\text{pure}}$ with $\mathcal{D}_{\text{pure}}$.

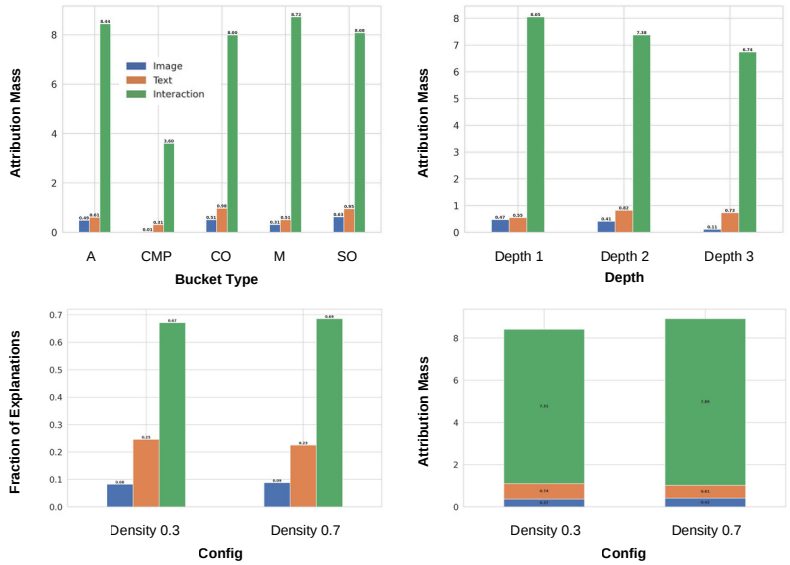

Fig. 12: EMAP attribution mass and explanation fractions for $\mathcal{M}_{\text{spur}}$ evaluated on $\mathcal{D}_{\text{spur}}$, segmented by query type, relational depth, and grid density.

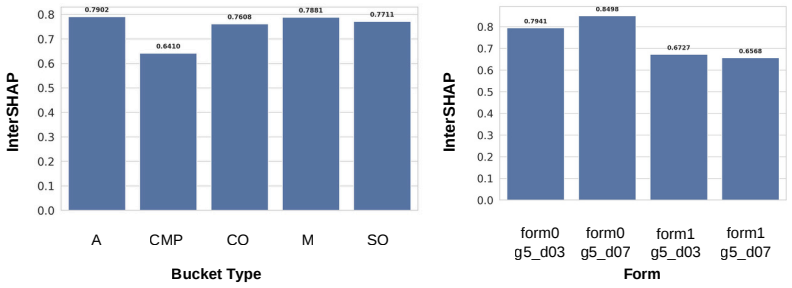

Fig. 13: InterSHAP score from $\mathcal{M}_{\text{spur}}$ with $\mathcal{D}_{\text{spur}}$.