

DiaVLo: Diagnosing Behaviours of Vision-Language Models

Lorenzo Corti*

Delft University of Technology
l.corti@tudelft.nl

Jie Yang

Delft University of Technology
j.yang-3@tudelft.nl

Abstract

Vision-language models (VLMs) rely on storing and transferring appropriate information across their sub-components. Verifying that the VLMs exhibit desired behaviours, while avoiding harmful ones, is central to their reliable deployment. Yet, methods that identify VLM behaviours remain scarce. We present DiaVLo, a diagnostic framework that leverages human curation and VLMs' generation capabilities to construct specifications of desired and observed VLM behaviours, surfacing potential misalignments. Beyond this, DiaVLo also provides causal estimates to identify the most influential concepts steering VLM behaviours. We evaluate DiaVLo on several open-source VLMs under both classification and generation conditions. Our experiments show that DiaVLo produces behaviour labels that correlate with model performance and provide context for measured performance. DiaVLo surfaced behaviours that are clearly aligned and misaligned, alongside patterns in how VLMs perceive, organise, and prioritise concepts.

1 Introduction

The success of current vision-language models (VLMs) hinges on the interplay of three primary modules: a visual encoder (e.g., CLIP (Radford et al., 2021)), a projection middle-layer, and a decoder-only LLM (e.g., Vicuna (Chiang et al., 2023)). The combination of these three elements, either composite (Liu et al., 2024b) or trained end-to-end (Bai et al., 2025), has proven effective across multimodal tasks in both offline benchmarks and downstream applications (Hao et al., 2022; Li et al., 2025). However, VLMs' performance appears to stem from the language module, as recent work has uncovered critical blind spots in visual perception (Tong et al., 2024) and performance discrepancies on vision-centric tasks (Fu et al., 2025). These findings point to unpredictable, under-specified VLM

* Corresponding author.

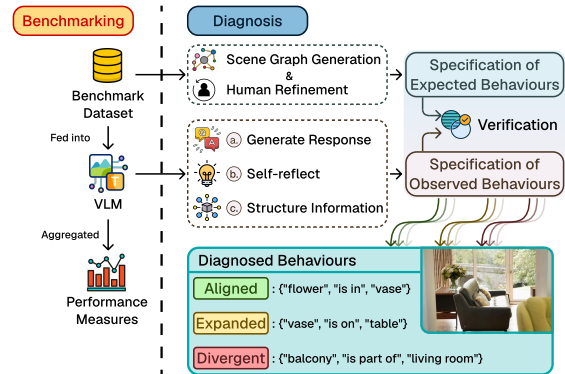


Figure 1: With DiaVLo, we frame VLM diagnosis as an addition to benchmarking, enabling the definition and verification of observed model behaviours with respect to specifications of desired behaviours.

behaviours that hinder our understanding of these models and their broader applicability across domains (Shu et al., 2025).

In this work, we address the problem of *diagnosing VLM behaviours*, i.e., verifying that these models exhibit the desired output regularities while avoiding unwanted ones (Holtzman et al., 2025; Zhang et al., 2022b; Cabrera et al., 2023). Prior work sought to analyse VLMs through *benchmarking* and *interpretability*. On the one hand, benchmarks provide aggregated assessments of VLMs' task performance, e.g., using embedding-based metrics (Zhang et al., 2020) or end-to-end ones (Liu et al., 2023b). Here, VLMs are commonly evaluated on diverse, human-interpretable capabilities such as visual perception and reasoning. For instance, SEED-Bench-2, organises 24k datapoints across 27 dimensions (Li et al., 2024a). On the other hand, interpretability methods (e.g., Achibat et al. (2024); Parcalabescu and Frank (2024)) describe salient visual and text features that VLMs may use to produce a response. These are generally classified as black-box methods (i.e., requiring query access to the model) or white-box methods (i.e., requiring full model access) (Resck et al.,

2025). Yet, these crucial works often rely on proxies that do not naturally map to real-world settings where humans might be actively using VLMs (Xiao et al., 2023; Sokol and Flach, 2020).

Given these limitations, we introduce DiaVLo, a diagnostic framework enabling the definition of desired behaviours and the extraction and classification of observed VLM behaviours in response to the former (Figure 1). First, desired behaviours are defined by creating symbolic representations of visual inputs and having humans verify their correctness and relevance to the task. Second, VLMs are prompted to verbalise the structured rationales underlying their responses, thereby producing self-explanatory representations of their own behaviours. Finally, desired and observed behaviours are semantically compared to identify behavioural alignment. In DiaVLo, we leverage causal modelling to quantify the influence of visual concepts on VLM outputs and mitigate potential inconsistencies in self-explained behaviours. For this, we synthetically generate counterfactual observed behaviours and estimate concept-level effects with a Double Machine Learning estimator (Chernozhukov et al., 2018).

We evaluate DiaVLo across four open-source VLMs and four public visual question answering datasets, in both classification and generation settings. Our results show that the behaviour labels assigned by DiaVLo correlate with measured model performance, indicating that DiaVLo can provide context to standard evaluation metrics. Additionally, DiaVLo surfaced behaviours that are clearly aligned and misaligned, while allowing us to uncover patterns related to (1) how VLMs perceive concepts (through hypernyms and hyponyms), (2) differences compared to humans in relating concepts together, and (3) prioritisation of certain concepts when responding to users’ requests. In summary, we contribute: a conceptualisation of the problem of diagnosing vision-language models; DiaVLo, a framework for surfacing and classifying the desired and observed VLM behaviours; and an extensive analysis of the behaviours of four well-known VLMs surfaced with DiaVLo.

2 Related Work

2.1 Vision-Language Models

Most of the current VLMs are image-and-text-to-text models that combine a visual encoder, a connection module between visual and text modalities,

and a decoder-only LLM. For the visual encoder, CLIP (Radford et al., 2021) is widely adopted (Laurençon et al., 2024). As connection modules, prior work used linear layers (Liu et al., 2023a), cross-attention (Alayrac et al., 2022), or ad-hoc methods (Li et al., 2023). Lastly, an LLM (e.g., Vicuna (Chiang et al., 2023)) handles text generation. Efforts around visual instruction tuning data improved visual understanding and grounding (Liu et al., 2024b; Peng et al., 2023).

Benchmarking VLMs Several benchmarks have been proposed to assess the capabilities of VLMs. Several benchmarks have been proposed to assess the capabilities of VLMs, *inter alia*, on visual question answering (Liu et al., 2023a; Goyal et al., 2017; Li et al., 2024a), low-level visual perception (Wu et al., 2024), visual relation grounding (Lu et al., 2025), and causal relevance graphs (Pratama et al., 2026). Despite their utility, benchmarks do not necessarily help identify or uncover specific behaviours (Holtzman et al., 2025) or study information transfer across modalities (Basu et al., 2024).

2.2 Explainable AI & Causality

A plethora of Explainable AI (XAI) methods have been proposed to produce *local* (sample-level) or *global* (class-level) output explanations, either in post-hoc (i.e., without altering a model) or self-explaining (i.e., embedded within a model) fashions (Zhang et al., 2021; Resck et al., 2025). Common XAI methods report the importance of individual input features (Selvaraju et al., 2017), identify influential or prototypical training samples (Koh and Liang, 2017; Chen et al., 2019), or produce concept-based (Balayn et al., 2021), rule-based (Ribeiro et al., 2018), or counterfactual explanations (Wachter et al., 2017). Notably, counterfactual explanations are the main category of explanations that incorporate causality despite calls for its broader adoption in XAI research (Guidotti, 2022).

Self-explanations Following advances in LLMs, natural language explanations can be easily produced alongside model outputs (Huang et al., 2023) to elicit reasoning and potentially surfacing reasoning traces (Wei et al., 2022). Recent work has found these explanations to be coherent and comparable to those from other XAI methods (Huang et al., 2023). In parallel, we are also starting to see LLM-specific adaptations of explanation evaluation criteria such as fidelity and self-consistency (Madsen et al., 2024; Parcalabescu and Frank, 2024).

2.3 Model Diagnosis

Model diagnosis is the process of investigating issues in a model and entails analysing symptoms, formulating hypotheses, and possibly reproducing the issue. It precedes *model debugging*, which focuses on ‘how’ to resolve issues. Model diagnosis is not trivial, as issues can stem from the data, the training process, or the model, *inter alia*. Prior work proposed diagnostic frameworks (Ribeiro et al., 2020; Ye et al., 2024), knowledge probes (Jawahar et al., 2019; May et al., 2019; Jiang et al., 2020), and techniques for mechanistic analysis (Bricken et al., 2023). Yet, findings from these (fundamental) works do not predictably transfer to observable model behaviours and their analysis (Holtzman et al., 2025; Sharkey et al., 2025).

Explanation-based Approaches Research has also suggested principled approaches to using XAI methods to verify model outputs and identify errors (Kulesza et al., 2015; Caruana et al., 2015; Han and Ghosh, 2021). The works by Balayn et al. (2021) and Sharifi Noorian et al. (2022) are of particular interest to ours as they seek to define desirable and observable model behaviours, respectively. We point to Lertvittayakumjorn and Toni (2021) for a comprehensive survey of these approaches.

3 Defining the VLM Diagnosis Problem

We define the VLM diagnosis problem as the process of analysing the associations that VLMs make across visual and textual inputs to produce their outputs. This requires one to construct (or access, if available) specifications of *desired* and *observed* behaviours. Note that, while observed behaviours can be extracted in various ways (subsection 2.2), desired behaviours often stem from domain requirements (Chen and Zhang, 2025) or human expectations (Lucy et al., 2024) and may entail additional effort in formulating. Hereafter, we use SHOULD-KNOW (SK; Balayn et al. (2021)) and REALLY-KNOW (RK; Sharifi Noorian et al. (2022)) for desired and observed behaviours, respectively.

We denote $\mathcal{D} = \{(x_0, Y_0), \dots, (x_{|\mathcal{D}|}, Y_{|\mathcal{D}|})\}$ as the set of a image-text pairs x_i and their ground truths Y_i , and a VLM $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{V}$, producing text token sequences $\hat{Y} = [\hat{y}_0, \dots, \hat{y}_m]$ where $\hat{y}_i \in \mathcal{V}$. To diagnose $\mathcal{M}$, we use SHOULD-KNOW and REALLY-KNOW behavioural specifications $s_i = (c_i, r_{ij}, c_j) \rightarrow \hat{Y}$ which relate human-understandable concepts $c_i \in \mathcal{C}$ through

r_{ij} used by $\mathcal{M}$ to produce $\hat{Y}$. We opt for a triplet-based definition to provide a structured, tractable representation of visual inputs. In general, SK specifications are task-specific but model agnostic, while RK specifications are specific to both. With these, we construct the following behaviour types:

- **Aligned:** $RK \subseteq SK$ – All of the behaviours exhibited by $\mathcal{M}$ meet the desired ones. More specifically, if $RK = SK$, we say the behaviours are *fully aligned*. Otherwise, if $RK \subset SK$, we say the behaviours are *partially aligned*.
- **Expanded:** $RK \cap SK \neq \emptyset \wedge \exists s_i^{RK} \notin SK$ – A VLM $\mathcal{M}$ partially exhibits the desired behaviours, but additional, unknown behaviours are observed.
- **Divergent:** $RK \cap SK = \emptyset$ – A VLM $\mathcal{M}$ does not exhibit any of the desired behaviours.

These behaviour definitions can be used to diagnose VLMs at the task level (e.g., visual question answering) to construct articulate model behaviours. While broadly applicable, we foresee the need to refine these definitions to better cover nuanced VLM behaviours.

4 The DiaVLo Framework

We illustrate the DiaVLo diagnostic framework (Figure 2) and its components: the definition of SHOULD-KNOWs, the extraction of REALLY-KNOWs, and the classification of VLM behaviours.

4.1 Defining SHOULD-KNOWs

The first step in DiaVLo is to construct SK specifications to serve as a diagnostic reference throughout DiaVLo to identify model behaviours and estimate the concept-level causal effects on VLM outputs. Concretely, we first use scene graph generation (SGG) to obtain a structured representation of images. Since these descriptions can be imprecise and generated in a task-agnostic manner (i.e., they do not account for the associated input prompt), we verify and expand them through human annotation to ensure correctness and task relevance.

4.1.1 Scene Graph Generation

Given an image, SGG produces: (1) a set of concepts $\mathcal{C} = \{c_i\} = \{(b_i, l_i)\}$ where each object is localised within the image with a bounding box b_i and belongs to a specific class l_i (e.g., person); and, (2) a set of relational triplets $\mathcal{E} = (c_i, r_{ij}, c_j)$ where r_{ij} relates concepts c_i and c_j (e.g., ‘held by’). In

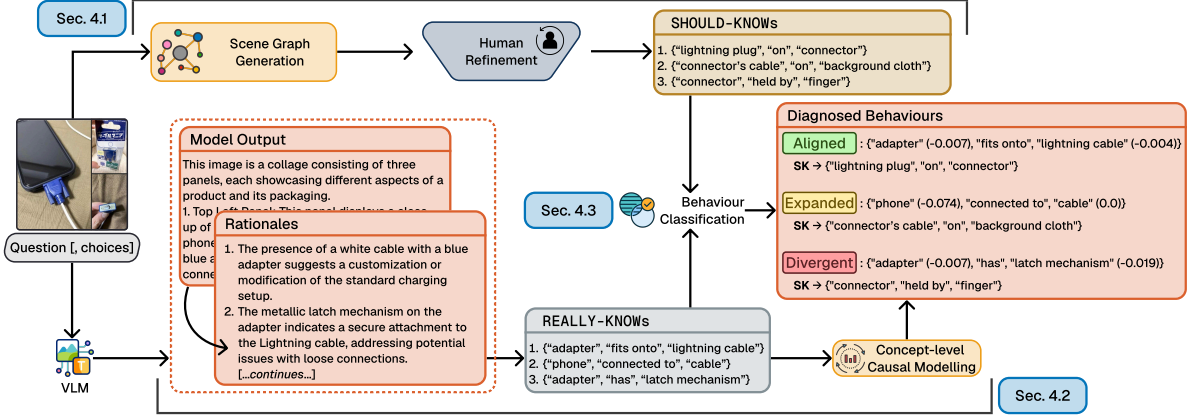


Figure 2: In DiaVLo, two parallel processes enable the classification of VLM behaviours. (Top) Scene graph generation and human curation produce SKs that are relevant for the visual and language inputs. (Bottom) VLMs provide RKs by verbalising rationales for their responses. We also compute concept-level causal estimates for deeper insights. (Middle) Finally, DiaVLo classifies model behaviours based on lexical and semantic matching.

DiaVLo, we use IETrans (Zhang et al., 2022a), a state-of-the-art method for unbiased SGG, which uses 70k distinct concept labels and 1.8k relationship labels. IETrans offers competitive performance, is lightweight, and provides long-tail coverage, avoiding collapsing to uninformative labels generated by other SGG methods. Note that while modern VLMs demonstrate open-vocabulary object detection capabilities (Li et al., 2024b; Wang and Liu, 2024), in our early tests, we found them noisier and less reliable than IETrans. Therefore, we did not pursue an empirical investigation of the use of VLMs in DiaVLo. This would also introduce a form of evaluation circularity and model self-preference bias (Panickssery et al., 2024).

4.1.2 Human Curation

We rely on human contributors to curate the descriptions obtained through SGG, since these can be imprecise and are agnostic to the associated text input. Despite the scalability of LLM-as-a-judge strategies, we opted for human curation, as LLM judges often require human-written references to combat performance degradation (Krumdick et al., 2025) and can exhibit documented preferential biases (Ye et al., 2025). Therefore, we leave the study of LLM judges for diagnosis as future work.

For DiaVLo, we design and implement two separate human curation tasks to verify and expand the SGG step’s output. In the **verification task**, contributors check the correctness and the task relevance (i.e., with respect to the input text and image) of the triplets extracted in the SGG step. Here, we assist them in copy-editing the triplets with a lightweight auto-complete mechanism based on the

concept and relationship labels used by IETrans. Note that contributors are not restricted to these label sets and can provide new ones already at this stage. Following this, with the **expansion task**, we seek to obtain additional triplets not identified in the SGG step. We asked contributors to indicate any relevant but missing triplets needed to extend the previously verified SK specifications, if any. Here, annotations comprise concept and relationship labels, as well as concept bounding boxes.

4.2 Extracting REALLY-KNOWs

We compose REALLY-KNOW specifications by leveraging self-explanations from VLMs (Madsen et al., 2024). While self-explanations (and chain-of-thought traces) may not always be faithful to the model (Madsen et al., 2024; Agarwal et al., 2024), they may be comparably plausible to other explanation approaches, particularly when a task is decomposed adequately (Huang et al., 2023; Turpin et al., 2023). In DiaVLo, we decompose the task into three steps, allowing the model to carry it out effectively (Wang et al., 2023; Randl et al., 2025). First, we collect the VLM’s responses to image-text pairs. Second, we prompt the VLM to produce rationales that corroborate its responses. Finally, we ask the VLM to format its rationales, for comparison with the SHOULD-KNOW specifications.

Localising Concepts We recover the bounding boxes of the RK concepts by first attempting to match RK and SK triplets. If unsuccessful, we use the open-vocabulary object detection model OWLv2 (Minderer et al., 2023) and similarity

matching as fall-back methods.¹ This information is also used in our causal modelling approach.

4.2.1 Causal Modelling of Concept Roles

While this procedure is lightweight, a zero-shot approach does not differentiate between the concepts VLMs use *vs* those they simply see. To quantify the role of different concepts, we leverage causal modelling to estimate their contributions to the observed VLM outputs. Note that, while causal modelling provides a verification layer to self-explanations, this differs from quantifying the fidelity of the RKs, which we leave for future work. Concretely, we model causal relationships in two steps: (1) transparent generation of counterfactual model input-output pairs, and (2) fitting of a double-machine learning (DML) estimator to obtain the concepts' causal effects.

Counterfactual Generation To quantify the effect that individual concepts have on model outputs, we first need counterfactual input-output pairs, with alternative concept compositions, and then observe the corresponding model outputs. Contrary to prior work that leverages neural methods (Alvarez-Melis and Jaakkola, 2017; Xu et al., 2021), we define a transparent, concept-level image perturbation strategy, keeping the input text constant.

For each *original* data sample, we take its $\text{RK}_{\mathcal{M}}^{x_i}$ specifications and compute the powerset $\mathbb{P}(\text{RK}_{\mathcal{M}}^{x_i})$. Each element $\mathcal{C} \in \mathbb{P}(\text{RK}_{\mathcal{M}}^{x_i})$ represents a combination of concepts $\{c_i, \dots, c_j\} \in \text{RK}_{\mathcal{M}}^{x_i}$. Based on these combinations, we use the bounding boxes of the concepts to create occlusion masks on the image, thereby obtaining *counterfactual inputs*. We prevent unwanted interventions on concepts by occluding solely the concepts within a given $\mathcal{C} \in \mathbb{P}(\text{RK}_{\mathcal{M}}^{x_i})$, leaving the others visible. Finally, we generate *counterfactual responses* $\hat{Y}^c$ by simply running inference with the VLM being inspected.

Causal Effect Estimation Given the inherent nonlinearity of VLMs, we opt for a DML approach (Chernozhukov et al., 2018). DML estimators do not rely on parametric assumptions (unlike, e.g., Gaussian mixture density estimators) and are better suited to capture complex relationships in data. For this, we define a set of covariates $Z_i = \mathcal{C}_{x_i} \setminus c_i$, the treatment variable $T_i = c_i$, and the outcome variable $\hat{Y}_{x_i} = \hat{Y}_{x_i}^c \cup \hat{Y}_{x_i}$. We then train two ML models: $\hat{Y}_{x_i} \approx \hat{f}(Z_i)$, approximating the outcome

given Z_{x_i} , and $T_i \approx \hat{g}(Z_i)$, approximating the treatment given Z_{x_i} . Finally, causal effects $\hat{\theta}_{c_i}$ are computed based on residuals $\hat{U}_i$ and $\hat{V}_i$, as:

$$\hat{U}_i = \hat{Y}_{x_i} - \hat{f}(Z_i) \quad (1) \quad \hat{V}_i = \mathcal{C}_{x_i} - \hat{g}(Z_i) \quad (2)$$

$$\hat{\theta}_{c_i} = \left(\frac{1}{n} \sum_{i=1}^n \hat{V}_i \mathcal{C}_{x_i} \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^n \hat{V}_i \hat{U}_i \quad (3)$$

The estimates $\hat{\theta}_{c_i}$ are, in practice, averages of multiple estimates obtained through cross-fitting. The models $\hat{f}$ and $\hat{g}$ are fitted on $(\phi(T_i, Z_i), \hat{Y}_{x_i})$ where $\phi(T, Z)$ encodes the presence (or absence) of the treatment variable and covariates (i.e., binary vector) based on the combinations in $\mathbb{P}(\text{RK}_{\mathcal{M}}^{x_i})$. In DiaVLo, we use gradient boosted trees.² See subsection D.2 for the implementation details.

Generally, causal estimators rely on a causal graph, i.e., a direct acyclic graph describing relationships between variables. In DiaVLo, we use the $\text{RK}_{\mathcal{M}}$ specifications as proxies for causal graphs. We discuss related caveats later in the paper.

Handling Longer Responses In DiaVLo, we extract the concepts $\{c_i, \dots, c_j\}$ from the original model response $\hat{Y}$ and construct binary vectors encoding the presence or absence of c_i in $\hat{Y}_{x_i}$. Then, we estimate the causal effects $|\{c_i, \dots, c_j\}|$ times given the Z_i covariates, as described above.

4.3 Classification of Behaviours

We propose classifying VLM behaviours from a semantic perspective, thereby accounting for naturally occurring language similarities. In DiaVLo, we quantify the cosine similarity $S_c(e(\text{SK}), e(\text{RK}))$ between the embedded SK and RK specifications.³ Given the SK and RK behaviours of a data sample, DiaVLo iteratively matches the RK with the SK with the largest S_c until either set is exhausted. This allows us to isolate both expected but unobserved behaviours and new ones. Behaviours can then be classified with flexible thresholding as:

- **Aligned:** if $|\text{RK}| \leq |\text{SK}|$ and $S_c \geq \tau_a$
- **Expanded:** if $|\text{RK}| > |\text{SK}|$ and $S_c \geq \tau_e$
- **Divergent:** $S_c < \tau_d$

²Theoretically, any ML model can be used for $\hat{f}$ and $\hat{g}$.

³We use `sentence-transformers/all-mpnet-base-v2`. The `ibm-granite/granite-embedding-english-r2` used before would erroneously assign $S_c > 0.8$ to all SK-RK pairs.

¹We use `ibm-granite/granite-embedding-english-r2`.

5 Experiments

We seek to answer two main questions: **(Q1)** What common behaviour types are identified in VLMs with DiaVLo?; and, **(Q2)** How informative are the behaviours found with DiaVLo?

Models & Datasets We apply DiaVLo on four well-known VLMs: InternVL2 8B (Chen et al., 2024), LLaVa-1.6 7B (Liu et al., 2024b), Qwen2.5-VL 7B (Bai et al., 2025), and ShareGPT4V 7B (Chen et al., 2023). We test these on four datasets:

- *LLaVa-Bench* (Liu et al., 2023a): All 60 open-ended, diverse, and challenging questions. We anticipate VLMs to struggle with this data.
- *MMBench* (Liu et al., 2024c): Repurposed captioning questions from the “Image Scene” (75) and “Image Topic” (64) classes as open-ended.
- *SEED-Bench 2* (Li et al., 2024a): Multiple-choice questions from the scene understanding (75) and visual reasoning (75) categories.
- *VQA v2* (Goyal et al., 2017): Multiple-choice questions from the “How many people are...” (75) and “What is the person...” (75) categories.

Crowdsourcing We recruited 520 workers (avg. wage: 8 GBP/h) on Prolific from English-speaking countries, with an approval rate of $\geq 95\%$. This study received ethics approval from our institution (ID: 4696). Refer to Appendix B for additional details and screenshots of the task interfaces.

Prompt Templates For each model-dataset combination, we fine-tune our prompts to ensure consistent instruction following. See Appendix C.

Implementation Details We self-host the VLMs on two NVIDIA A10 GPUs. We use their default parameterisation, limit response length to 256 tokens, and employ greedy decoding for reproducibility. Additional details in Appendix D.

(Q1) Common Model Behaviour Types We use DiaVLo to classify VLM behaviours as Aligned, Expanded, and Divergent as proposed in subsection 4.3. To this end, we also report on our exploration and selection of semantic similarity thresholds. Finally, we qualitatively inspect 200 samples (covering 952 RKs) to surface recurring behaviours.

	LLaVa-Bench	MMBench	SEED-Bench 2	VQA v2
Initial	580 (-)	1442 (-)	1783 (-)	1798 (-)
Validation	300 (51.7%)	839 (58.2%)	1042 (58.4%)	1051 (58.5%)
Expansion	516 (43.4%)	1024 (42.1%)	1233 (38.7%)	1302 (46.0%)
Expert	546 (11.0%)	1094 (5.9%)	1301 (8.4%)	1391 (7.8%)
M_{SK} (IQR)	9 (5.75, 11)	7 (5, 10.5)	9 (5, 12)	8.5 (6, 12)

Table 1: SKs collected and percentages of relevant SKs (Validation), of new SKs (Expansion), or of expert edits (including median (M_{SK}) and inter-quartile range (IQR)).

(Q2) Informativeness of Model Behaviours We study the informativeness of behaviours from two perspectives. First, we study how well the similarities between the SK and RK produced by DiaVLo can describe model performance. Even if a VLM performs satisfactorily, it is not a given that the behaviours DiaVLo finds align with performance. This last point would likely require a mechanistic analysis, which is outside the scope of this work. Since we cannot assume a priori that these two variables are linearly related, we bootstrap Mutual Information (MI).⁴ Second, we do model-pairwise comparisons across datasets to explore the distributions of concept-level causal effects. To mitigate potential biases, we report the median causal estimates across $n = 100$ random data splits.

6 Results & Discussion

6.1 Overview of SK and RK specifications

Inspecting SHOULD-KNOWS Each human annotator worked on 5 data samples, covering a varying number of candidate SKs. Each sample was shown to a single human annotator.⁵ Afterwards, with the help of colleagues from our computer science department, we manually verified that the crowdsourced SKs behaviours (Table 1) were properly formatted and correctly aligned with the input text and image. We found that a maximum of $\approx 11\%$ (Table 1; Expert row) of behaviours required corrections for trivial mistakes – mostly typos and incorrectly formatted data. In a few cases, contributors provided valid triples even when only the concept or relation labels were expected, thereby slightly increasing the final number of SKs.

Inspecting REALLY-KNOWS We extracted from the four VLMs RKs for 1754/1996 initial samples, los-

⁴Refer to subsection E.4 for our MI bootstrapping procedure, including significance testing and comparison with randomly shuffled behaviour similarities.

⁵See subsection E.1 for ablations on IETrans. We also note that, by first showing contributors triplets produced with IETrans, they may have exhibited anchoring bias.

	LLaVa-Bench		MMBench		SEED-Bench 2		VQA v2	
	N_s (N_{RK})	M_{RK} (IQR)	N_s (N_{RK})	M_{RK} (IQR)	N_s (N_{RK})	M_{RK} (IQR)	N_s (N_{RK})	M_{RK} (IQR)
InternVL2	58 (565)	10 (7, 12)	139 (1287)	9 (7, 11)	150 (1254)	8 (6, 10)	150 (1014)	7 (4, 9)
LLaVa-1.6	44 (279)	7 (3, 8.25)	133 (823)	6 (4, 8)	129 (853)	6 (4, 10)	142 (939)	6 (4, 9)
Qwen2.5-VL	56 (504)	9 (6, 12)	139 (601)	4 (3, 5)	150 (554)	3 (2, 4)	150 (336)	2 (2, 2)
ShareGPT4V	49 (279)	5 (3, 8)	135 (645)	4 (2, 7)	109 (241)	1 (1, 2)	21 (46)	1 (1, 1)

Table 2: RK statistics shown across models and datasets: raw number of RK extracted (N), median (M), and inter-quartile range (IQR). We highlight model-dataset pairs based on the number of RK successfully extracted: $> 90\%$ in green, $70 - 90\%$ in yellow, and $< 70\%$ in red. We include results from ShareGPT4V on VQA v2 for transparency.

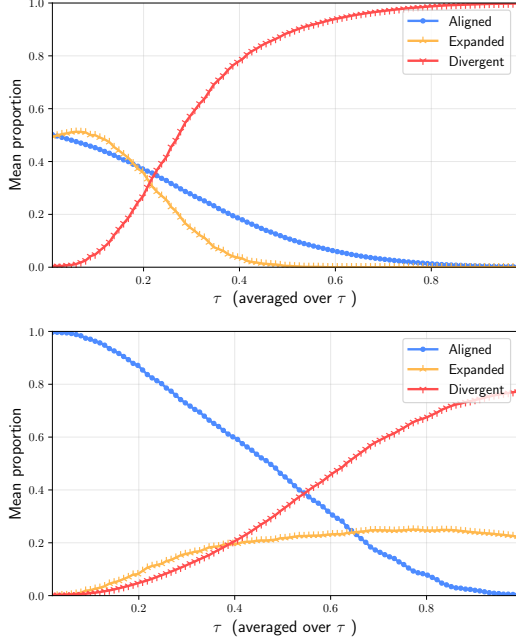


Figure 3: Behaviour thresholds’ sensitivity to τ_a (top) and τ_e (bottom). τ_d is set afterwards, based on τ_e . Additional visualisations are included in subsection E.2.

ing $\approx 12\%$ due to empty responses and unstable instruction-following, with ShareGPT4V as the loss leader (Table 2). We cleaned up the repeated RKs produced by the VLMs (likely due to greedy decoding) by coalescing them into unique ones.

6.2 Common Model Behaviour Types (Q1)

Setting Behaviour Thresholds We perform a sensitivity analysis to set the behaviour thresholds τ_a , τ_e , and τ_d (Figure 3). We found that τ_a plateaus in $[0.65, 0.8]$, while a $\tau_e \geq 0.4$ leads to an increasing number of incorrectly-classified, Divergent samples. To corroborate this, we manually inspected 300 aligned SK-RK pairs and 300 divergent SK-RK pairs.⁶ With $\tau_a \geq 0.7$, we found 22.31% of the pairs had only one similar concept and were

falsely classified as aligned. With a more conservative $\tau_a \geq 0.75$, we reduced this to 17.70% so that cases with one similar concept and relationship are accounted for. This also increased the average similarity $\overline{S_c}$ from 0.77 to 0.84. On the other hand, setting $\tau_d < 0.35$ already led to 92.69% of the pairs having dissimilar or distinct concepts and relationships ($\overline{S_c} = 0.21$). In DiaVLo, we thus use: $\tau_a \geq 0.75$; $\tau_e \geq 0.35$; and $\tau_d < 0.35$.

Behaviours Identified Figure 4 shows the distribution of behaviours classified by DiaVLo. Overall, we observe that the VLMs we tested mostly exhibit expanded or divergent behaviours. Refer to subsection E.5 for complete behaviour examples.

(1) Aligned Behaviours (6.4%): DiaVLo highlighted a low number of aligned behaviours in the VLMs tested. Here, the RKs are highly similar to the corresponding SKs: concept and relationship labels are often identical and structured in the same way. The main qualitative difference for aligned behaviours concerns the use of active or passive voice when formulating a SK or a RK, such as:

Qwen2.5-VL on SEED-Bench 2 (Sample ID: 57)

RK: {"guitar", "being played by", "person"}
 $\hookrightarrow$ **SK_{match}:** {"player", "playing", "guitar"}

Predictably, the SK and RK from LLaVa-Bench differ significantly despite comparable performance.

(2) Expanded Behaviours (52.5%): Overall, the expanded behaviours cover SK-RK pairs that are thematically related (e.g., clothing) and exhibit semantic or structural differences:

1. *Describing over Composing:* We found RKs to be largely descriptive (e.g., the properties of objects) rather than compositional (e.g., spatial locations of objects). This pattern aligns with prior research of VLMs’ blind spots and gaps in their visual capabilities (Fu et al., 2025).

⁶ $S_c: \wedge = 0.021; \vee = 0.92; \overline{S_c} = 0.506; \sigma_{S_c} = 0.152$.

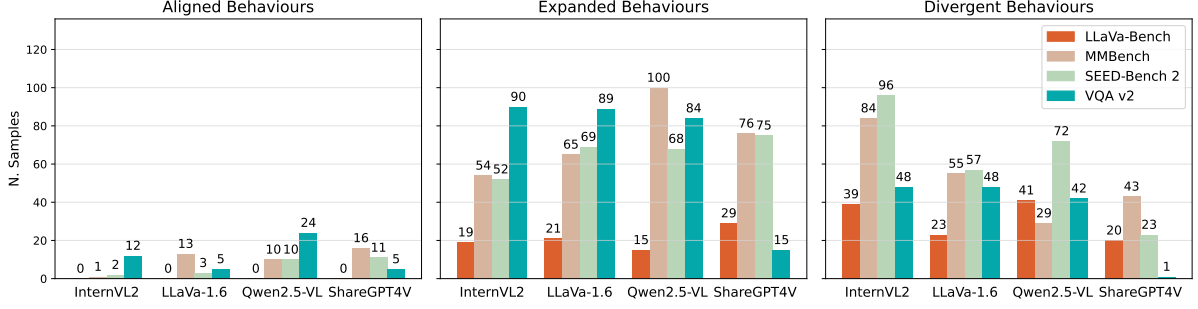


Figure 4: Identified Model Behaviours.

2. *Organisation of Concepts*: While similar (or identical) concepts are covered in both SKs and RKs, the relationships and concept combinations differ. This might hint at concept-level, latent prioritizations applied by VLMs that, under the same circumstances and input data, do not align with those of humans. We note that these do not necessarily constitute faulty behaviours, but should be identified and assessed nonetheless.

InternVL2 on VQAv2 (Sample ID: 26130006)

RK: {"person", "near", "wave"}
 ↳SK_{match}: {"man", "going for", "wave"}

RK: {"surfboard", "part of", "person"}
 ↳SK_{match}: {"man", "riding", "surfboard"}

(3) **Divergent Behaviours (41.1%)**: Here, we have cases in which the VLMs rely on very different concepts and relations compared to the SKs, or organise those concepts in very different ways, more so than Expanded behaviours. Interestingly, divergent behaviours are also those with a larger number of RKs. Here, the VLMs may be more uncertain about their responses and may attempt to make their output self-consistent. Predictably, the SKs and RKs related to LLaVa-Bench differ significantly despite comparable measured performance.

ShareGPT4V on LLaVa-Bench (Sample ID: 33)

RK: {"mug", "white", "ceramic"}
 ↳SK_{match}: {"letter", "written on", "mug"}

RK: {"mug", "slightly tilted", "left"}
 ↳SK_{match}: {"cap", "covering", "head"}

6.3 Informativeness of Model Behaviours (Q2)

Performance-Behaviours Relationship Overall, SKs and RKs from DiaVLo are good indicators of VLM performance (Table 3). Note that their differing magnitudes are due to how performance is

VLM	LLaVa-Bench		MMBench		SEED-Bench 2		VQA v2	
	A	MI	A	MI	A	MI	A	MI
InternVL2	0.673	3.367	0.796	3.200	0.787	0.379	0.667	0.476
LLaVa-1.6	0.657	2.389	0.792	2.512	0.667	0.438	0.585	0.481
Qwen2.5-VL	0.670	3.008	0.743	1.561	0.833	0.188	0.700	0.147
ShareGPT4V	0.687	2.058	0.705	2.116	0.661	0.292	0.810	0.558 [†]

Table 3: Measured accuracy (A) vs Mutual Information $MI(A, S_c)$. [†]: Very small sample size: see Table 2.

computed. For LLaVa-Bench and MMBench, performance is continuous within $[0, 1]$, and MI is theoretically unbounded. For the other datasets, instead, performance is discrete in $\{0, 1\}$, and $MI \leq \ln(2) \simeq 0.693$. Model-wise, we found that InternVL2 and LLaVa-1.6 exhibit behaviours more closely related to their performance. Qwen2.5-VL behaved consistently only on LLaVa-Bench, while falling off on the rest. Finally, ShareGPT4V showed better consistency on open-ended data, likely due to its focus on image captioning.

Causal Effect of Concepts We find that the causal estimates produced by DiaVLo concentrate generally around 0.0, with selected spots offset from it (Figure 5). Specifically, estimates computed on open-ended datasets span multiple concepts and may disperse across the range of output tokens (Figure 5b). Instead, results from multiple-choice datasets show a higher concentration of zero-valued estimates, possibly indicating that VLMs activate on concepts irrelevant to specific questions (Figure 5c). These results indicate that DiaVLo identifies concepts that are more likely to influence specific answers. Note that, despite carefully occluding images to avoid interference across visual concepts, fine-grained masking approaches (Pawlowski et al., 2020; Melistas et al., 2024; Rasal et al., 2025) could cause VLMs to produce different counterfactual responses and, in turn, different

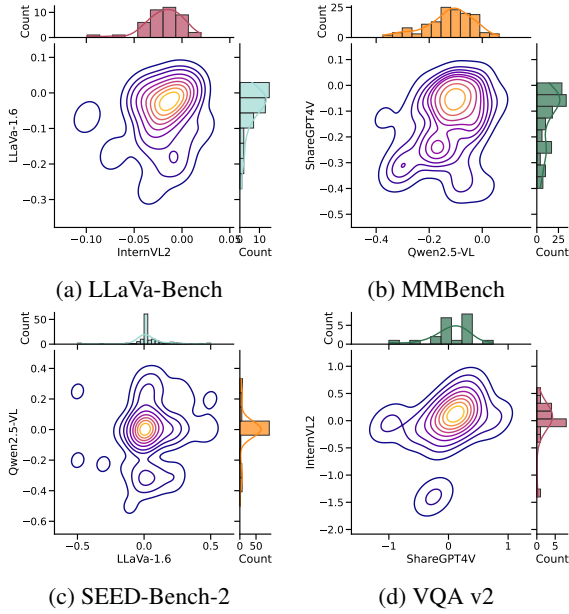


Figure 5: Example distributions of causal effects. Full comparisons for all VLMs and dataset combinations can be found in [subsection E.6](#).

causal estimates.⁷ Therefore, definitive conclusions will likely require triangulation with multi-modal interpretability research ([Liu et al., 2025](#)).

REALLY-KNOWs as Causal Graphs We checked the correctness of the RK graphs and the identifiability of the estimands. First, when creating the causal model from the graph structure, we used structural refutation tests and obtained mixed results. RKs were not falsified when sufficient counterfactual samples could be produced ($\gtrsim 100$ in our data), and the RKs were sufficiently informative and passed baseline permutation tests. Otherwise, we consider the results likely inconclusive due to low test power. Second, we identified estimands based on the assumed RK structures. Given the RK collected, this step did not surface issues. We ran additional placebo refutation tests on the estimation process, which were also inconclusive. Nonetheless, further research is required to fully characterise the use of REALLY-KNOWs as causal graphs.

7 Conclusion

We introduced DiaVLo, a framework to diagnose the behaviours of VLMs. By constructing specifications of expected and observed VLM behaviour, DiaVLo surfaces potential misalignments and quan-

tifies the effects of the concepts used by VLMs for inference. Beyond behaviours that are clearly aligned or misaligned, our results show that VLMs can exhibit behaviours that are more nuanced: VLMs can be selective in the concepts they use, and relate them differently from how humans do, potentially favouring broad descriptions over compositional aspects of visual inputs. Works like DiaVLo provide tools for identifying VLM behaviours, supporting inquiries into composite, system-level behaviours, and potentially informing mitigation techniques for curbing unwanted behaviours.

Limitations

Faithfulness of Self-explanations Our work relies on VLM-generated self-explanations. Despite adding a verification layer through causal modelling, this is not equivalent to assessing the faithfulness of the self-explanations produced by the VLMs. Even if less reliable than other explainable AI approaches in some cases ([Huang et al., 2023](#); [Randl et al., 2025](#)), these explanation methods are still susceptible to biases and spurious correlations. They may treat input modalities separately ([Kazmierczak et al., 2025](#)). Possible approaches to mitigate this, and improve the faithfulness of VLMs’ self-explanations, could use factored decomposition ([Radhakrishnan et al., 2023](#)) to avoid spurious information from the original question. Yet, to the best of our knowledge, these may still provide limited faithfulness for VLMs ([Li et al., 2026](#); [Lee et al., 2026](#); [Uppaal et al., 2026](#)). Nonetheless, as VLMs’ capabilities and limitations are actively researched ([Liu et al., 2024a](#); [Fu et al., 2025](#)), self-explanations offer a practical means of generating REALLY-KNOW specifications for VLMs.

Human Dependency and Scalability DiaVLo currently employs human curation to refine candidate SKs and ensure that they are relevant to the corresponding request. Yet, this dependency poses per-dataset scalability hurdles. While we believe in the growing need for human curation, more scalable SHOULD-KNOW curation pipelines could draw on established crowdsourcing workflows ([Grunde-McLaughlin et al., 2025](#)). For example, one could implement a Find-Fix-Verify workflow ([Bernstein et al., 2010](#)) leveraging the increasing capabilities of VLMs. First, in the *Find* step, SK candidates that are repeated, likely vague, incorrect, or likely irrelevant to the request are identified. Then, in the *Fix* step, probable issues can be assigned to

⁷See [subsection E.3](#) for ablations on OWLv2, which we used to locate visual concepts from RKs not found directly (lexical match) in the SK specifications.

another VLM or to humans, depending on their severity, following a Map-Reduce pattern (Kittur et al., 2011). Finally, in the *Verify* step, humans verify the previously applied fixes. The human dependency, in the average case, would therefore be restricted to a subset of the SKs, smaller than what we had contributors work on in DiaVLo. A similar pipeline would require balancing precision and recall at the *Find* step, to avoid missing mistakes, ensure that complex and nuanced behaviours are robustly assessed, and tune the router in the *Fix* step to keep the cost of querying VLM judges and human contributions under budget constraints.

Unstable Instruction-following We found that the VLMs tested struggled with the questions in the datasets we used, sometimes returning empty strings. This happened across different datasets and with different decoding strategies (greedy, beam search, and sampling). Similarly, the VLMs tested struggled to follow the given template when asked to structure their rationales. We suspect this shortcoming is a by-product of the training data (predominantly conversational) and the model size.

Not Testing on Larger Models In this work, we dealt with VLMs with parameter counts in the 7B-8B range. Larger VLMs (seem to) “fill in more blanks” compared to smaller ones, often resulting in stronger raw performance. Given our results and related work on VLM blindspots and idiosyncrasies between raw performance and model internals (Tong et al., 2024; Fu et al., 2025), we are inclined to believe that similar patterns will be exhibited by larger models as well, despite what leaderboards report. We acknowledge that applying our framework to larger models may yield conclusions different from those we obtained.

Not Testing on Closed-source Models Closed-source models, e.g., GPT-5, are intermittently updated without public notice. While it is important to analyse these models, given their widespread use, we refrained from using them in our experiments, as doing so would undermine the reproducibility of our results and conclusions.

Ethical Considerations

Perpetuating Harmful Behaviours While DiaVLo is meant to improve VLMs by helping identify misaligned behaviours, this information can also be used to further steer models in harmful directions. A malicious actor could use DiaVLo

to reduce appropriate behaviours and obtain a VLM that perpetuates stereotypical or hateful outputs. To the best of our knowledge, we did not encounter any harmful model outputs in our experiments. Bridging behavioural analysis of VLMs with red teaming and vulnerability disclosure practices could help mitigate the risk of misuse. Practitioners working in these areas have protocols for disclosing undesirable or harmful model behaviours to model creators and labs, for architecture- and provider-specific verifications, and for determining the level of model access needed. We believe that (1) a more coordinated approach would be needed for reporting concerning behaviours (e.g., as in Longpre et al. (2026)) and (2) research at the intersection of interpretability and model steering is crucial to build methods and toolkits to both understand (e.g., through interpretability) and act (e.g., through RL) on model behaviours. Other solutions may only provide temporary, model-specific patches (e.g., through developer-prompt instructions).

Participant Safety and Consent Our data collection to define SHOULD-KNOW specifications was conducted with ethics approval from our institution (ID: 4696). All participants from Prolific provided explicit consent before participation and could withdraw and have their contribution deleted without explanation or penalty. The datasets and classes in our experiments were screened to exclude potentially offensive data points. To the best of our knowledge, this was achieved, as no participants reported problems during the task.

Data Privacy All responses were anonymised, and no personally identifiable information was collected. Data will be released in processed form to ensure participant privacy.

References

- Reduan Achitibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. [AttnLRP: Attention-aware layer-wise relevance propagation for transformers](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR.
- Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. [Faithfulness vs. plausibility: On the \(un\)reliability of explanations from large language models](#). *Preprint*, arXiv:2402.04614.

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.
- David Alvarez-Melis and Tommi Jaakkola. 2017. [A causal framework for explaining the predictions of black-box sequence-to-sequence models](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Agathe Balayn, Panagiotis Soilis, Christoph Lof, Jie Yang, and Alessandro Bozzon. 2021. [What do you mean? interpreting image classification with crowd-sourced concept extraction and analysis](#). In *Proceedings of the Web Conference 2021*, WWW '21, page 1937–1948, New York, NY, USA. Association for Computing Machinery.
- Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024. [Understanding information storage and transfer in multi-modal large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 7400–7426. Curran Associates, Inc.
- Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. [Soylent: a word processor with a crowd inside](#). In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, page 313–322, New York, NY, USA. Association for Computing Machinery.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. [Improving human-ai collaboration with descriptions of ai behavior](#). *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).
- Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. [Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission](#). In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA. Association for Computing Machinery.
- Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. 2019. [This looks like that: Deep learning for interpretable image recognition](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. [Sharegpt4v: Improving large multi-modal models with better captions](#). *Preprint*, arXiv:2311.12793.
- Quan Ze Chen and Amy Xian Zhang. 2025. [Case law grounding: Using precedents to align decision-making for humans and ai](#). In *Proceedings of the ACM Collective Intelligence Conference*, CI '25, page 226–238, New York, NY, USA. Association for Computing Machinery.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. [Double/debiased machine learning for treatment and structural parameters](#). *The Econometrics Journal*, 21(1):C1–C68.
- Wei-Lin Chiang, Zhuohan Li, Zilong Lin, Ying Sheng, Zhonghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#). Accessed: 2024-04-22.
- David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

- Povilas Daniušis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. 2010. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 143–150, Arlington, Virginia, USA. AUAI Press.
- Josè A. R. Fonollosa. 2019. *Conditional Distribution Variability Measures for Causality Detection*, pages 339–347. Springer International Publishing, Cham.
- Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. [Hidden in plain sight: Vlms overlook their visual representations](#). In *Proceedings of the Second Conference on Language Modeling (COLM 2025)*.
- Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. [Review of causal discovery methods based on graphical models](#). *Frontiers in Genetics*, 10.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Madeleine Grunde-McLaughlin, Michelle S. Lam, Ranjay Krishna, Daniel S. Weld, and Jeffrey Heer. 2025. [Designing llm chains by adapting techniques from crowdsourcing workflows](#). *ACM Trans. Comput.-Hum. Interact.*, 32(3).
- Riccardo Guidotti. 2022. [Counterfactual explanations and how to find them: literature review and benchmarking](#). *Data Mining and Knowledge Discovery*.
- Xing Han and Joydeep Ghosh. 2021. [Model-agnostic explanations using minimal forcing subsets](#). In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. [Language models are general-purpose interfaces](#).
- Jennifer Hill and Elizabeth A. Stuart. 2015. [Causal inference: Overview](#). In James D. Wright, editor, *International Encyclopedia of the Social and Behavioral Sciences (Second Edition)*, second edition edition, pages 255–260. Elsevier, Oxford.
- Ari Holtzman, Peter West, and Luke Zettlemoyer. 2025. [Generative models as a complex systems science: How can we make sense of large language model behavior?](#) *Journal of Social Computing*, 6(2):75–94.
- Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. [Nonlinear causal discovery with additive noise models](#). In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.
- Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. [Can large language models explain themselves? a study of llm-generated self-explanations](#). *Preprint*, arXiv:2310.11207.
- Aapo Hyvärinen and Petteri Pajunen. 1999. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. [What does BERT learn about the structure of language?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, and Gianni Franchi. 2025. [Explainability and vision foundation models: A survey](#). *Information Fusion*, 122:103184.
- Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. [Crowdforge: crowdsourcing complex work](#). In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, page 43–52, New York, NY, USA. Association for Computing Machinery.
- Pang Wei Koh and Percy Liang. 2017. [Understanding black-box predictions via influence functions](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.
- Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. [No free labels: Limitations of llm-as-a-judge without human grounding](#). *Preprint*, arXiv:2503.05061.
- Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. [Principles of explanatory debugging to personalize interactive machine learning](#). In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, page 126–137, New York, NY, USA. Association for Computing Machinery.
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. [Building and better understanding vision-language models: insights and future directions](#). *Preprint*, arXiv:2408.12637.
- Eunsoo Lee, Jeongwoo Lee, Minki Hong, Jangho Choi, and Jihie Kim. 2026. [VisDoT : Enhancing visual reasoning through human-like interpretation grounding and decomposition of thought](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 610–640, Rabat, Morocco. Association for Computational Linguistics.

- Piyawat Lertvittayakumjorn and Francesca Toni. 2021. [Explanation-based human debugging of NLP models: A survey](#). *Transactions of the Association for Computational Linguistics*, 9:1508–1528.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. 2024b. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28076–28086.
- Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1587–1606.
- Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Lee Boyd-Graber, Haitao Mi, and Dong Yu. 2026. [Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization](#). In *The Fourteenth International Conference on Learning Representations*.
- Hanchao Liu, Wenyan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. [A survey on hallucination in large vision-language models](#). *Preprint*, arXiv:2402.00253.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yiming Liu, Yuhui Zhang, and Serena Yeung-Levy. 2025. [Mechanistic interpretability meets vision language models: Insights and limitations](#). In *The Fourth Blogpost Track at ICLR 2025*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. [Mmbench: Is your multi-modal model an all-around player?](#)
- Shayne Longpre, Elaine Zhu, Carson Ezzell, Avijit Ghosh, Sean McGregor, Kevin Paeth, Kevin Klyman, Sayash Kapoor, Rishi Bommasani, Ruth E. Appel, Gregory Strom, Lauren McIlvenny, Mark M. Jaycox, Peter Slattery, Nathan Butters, Arvind Narayanan, Percy Liang, and Alex Pentland. 2026. [FLARE-AI: Flaw reporting for AI](#). In *Forty-third International Conference on Machine Learning*.
- Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. 2025. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19618–19627.
- Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. 2024. [“one-size-fits-all”? examining expectations around what constitute “fair” or “good” NLG system behaviors](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1054–1089, Mexico City, Mexico. Association for Computational Linguistics.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. [Are self-explanations from large language models faithful?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. [On measuring social biases in sentence encoders](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.
- Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vrontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. 2024. [Benchmarking counterfactual image generation](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 133207–133230. Curran Associates, Inc.
- Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. [Scaling open-vocabulary object detection](#). In *Advances in Neural Information Processing Systems*,

- volume 36, pages 72983–73007. Curran Associates, Inc.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [Llm evaluators recognize and favor their own generations](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.
- Letitia Parcalabescu and Anette Frank. 2024. [On measuring faithfulness or self-consistency of natural language explanations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.
- Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. 2020. [Deep structural causal models for tractable counterfactual inference](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 857–869. Curran Associates, Inc.
- Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*, 1st edition. Basic Books, Inc., USA.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. [Kosmos-2: Grounding multimodal large language models to the world](#). *Preprint*, arXiv:2306.14824.
- Dhita Putri Pratama, Soyeon Caren Han, and Yihao Ding. 2026. [Diagnosing causal reasoning in vision-language models via structured relevance graphs](#). *Preprint*, arXiv:2602.20878.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuotė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. [Question decomposition improves the faithfulness of model-generated reasoning](#). *Preprint*, arXiv:2307.11768.
- Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2025. Evaluating the reliability of self-explanations in large language models. In *Discovery Science*, pages 36–51, Cham. Springer Nature Switzerland.
- Rajat R Rasal, Avinash Kori, Fabio De Sousa Ribeiro, Tian Xia, and Ben Glocker. 2025. [Diffusion counterfactual generation with semantic abduction](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 51201–51228. PMLR.
- Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. [Explainability and interpretability of multilingual large language models: A survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20454–20486, Suzhou, China. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. [Anchors: High-precision model-agnostic explanations](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2022. [What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition](#). In *Proceedings of the ACM Web Conference 2022, WWW '22*, page 882–892, New York, NY, USA. Association for Computing Machinery.
- Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. 2025. [Open problems in mechanistic interpretability](#). *Transactions on Machine Learning Research*. Survey Certification.
- Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. [Large vision-language model alignment and misalignment: A survey through the lens of explainability](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1713–1735, Suzhou, China. Association for Computational Linguistics.

- Kacper Sokol and Peter Flach. 2020. [Explainability fact sheets: A framework for systematic assessment of explainable approaches](#). In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 56–67, New York, NY, USA. Association for Computing Machinery.
- Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.
- Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. 2026. [Journey before destination: On the importance of visual faithfulness in slow thinking](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4147–4168, Rabat, Morocco. Association for Computational Linguistics.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. [Counterfactual explanations without opening the black box: Automated decisions and the gdpr](#). *Harvard Journal of Law & Technology (Harvard JOLT)*, 31(2):841–888.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.
- Yuxuan Wang and Xiaoyuan Liu. 2024. [Predicate debiasing in vision-language models integration for scene graph generation enhancement](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1627–1639, Miami, Florida, USA. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Naftali Weinberger. 2018. [Faithfulness, coordination and causal coincidences](#). *Erkenntnis*, 83(2):113–133.
- Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. [Q-bench: A benchmark for general-purpose foundation models on low-level vision](#). *Preprint*, arXiv:2309.14181.
- Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. [Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.
- Shuyuan Xu, Yunqi Li, Shuchang Liu, Zuohui Fu, Xu Chen, and Yongfeng Zhang. 2021. [Learning post-hoc causal explanations for recommendation](#). *Preprint*, arXiv:2006.16977.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. [Justice or prejudice? quantifying biases in LLM-as-a-judge](#). In *The Thirteenth International Conference on Learning Representations*.
- Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. [FLASK: Fine-grained language model evaluation based on alignment skill sets](#). In *The Twelfth International Conference on Learning Representations*.
- Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. 2022a. Fine-grained scene graph generation with data transfer. In *Computer Vision – ECCV 2022*, pages 409–424, Cham. Springer Nature Switzerland.
- Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2022b. [Machine learning testing: Survey, landscapes and horizons](#). *IEEE Transactions on Software Engineering*, 48(1):1–36.
- Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2011. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI'11*, page 804–813, Arlington, Virginia, USA. AUAI Press.
- Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. 2015. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020*,

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. [Explain and predict, and then predict again](#). In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21*, page 418–426, New York, NY, USA. Association for Computing Machinery.

A Primer on Causal Inference

Causal inference is the “discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data” (Hill and Stuart, 2015). Specifically, causal inferences aim to estimate the effect of one variable (e.g., input feature) on another (e.g., model prediction) (Pearl and Mackenzie, 2018). Randomised Control Trials (RCT) are commonly adopted to estimate such effects. In RCTs, two groups (e.g., groups of people) are administered a *treatment* to assess its *effect* – compared to not administering such a treatment – on the outcome of an experiment. RCTs are carried out abiding to the *ceteris paribus* principle (i.e., “all other things being equal”): the treatment is the only variable within a trial.

Unfortunately, carrying out RCTs and obtaining *counterfactual* data that answers to “what-if” questions (e.g., ‘What if we had not administered a medicine?’) can be expensive, infeasible, or unethical to be carried out. When access to counterfactual data is not possible, Causal Discovery can be leveraged to find causal relationships and form a so-called causal graph.

In the following, we describe these two concepts.

A.1 Causal Graphs

Causal graphs are modelling tools that show the relations and effects a set of (possibly) interrelated independent variables may have on the final outcome Y (i.e., the dependant variable) through a directed acyclic graph (DAG) (Pearl and Mackenzie, 2018). Causal graphs are particularly useful for studying different interventions (i.e., the treatments) without performing a real trial. Here, we briefly introduce terms that describe the role that individual variables can take within a causal graph.

Confounder: a variable Z which has an effect on other variables, e.g., X and Y , such that X and Y show correlation despite not being causally related. Confounders need to be accounted for

when studying the relationship between X and Y .

$$X \leftarrow Z \rightarrow Y \quad (4)$$

Mediator: a variable M causally related to an independent variable X causing an *indirect* effect on the outcome Y .

$$X \rightarrow M \rightarrow Y \quad (5)$$

Collider: a variable C that is influenced by two or more variables X and Y .

$$X \rightarrow C \leftarrow Y \quad (6)$$

When dealing with such variables, one is usually interested in estimating the *Average Treatment Effect* (ATE). That is, the average difference between administering a treatment and not administering it across the population being studied.

A.2 Causal Discovery

Generally, creating a causal graph requires domain expertise and is disentangled from the experimental hypotheses. Still, researchers have proposed techniques to infer causal structures from observational data by relying on statistical independence tests: These fall under the umbrella of Causal Discovery. Here, we report the main algorithms available. Refer to (Glymour et al., 2019) for a complete categorisation of causal discovery methods. First, constraint-based causal discovery algorithms, like Peter-Clark (PC) and Fast Causal Inference (FCI) (Spirtes et al., 2000), are based on a complete and undirected graph including all the variables involved and use statistical (conditional) independence tests to prune the edges. Second, score-based models like Greedy Equivalence Score (GES) (Chickering, 2002) start with an empty graph and add edges as long as the scoring function (e.g., Bayesian Information Criterion) increases. Then, edge removal operations are applied to test whether the score can be further increased. Finally, pairwise approaches aim to define causal relations between any two variables by evaluating the fitness of the data to an additive noise model (Hoyer et al., 2008), by bidirectionally comparing the standard deviation of the rescaled values of one variable with respect to the other one in the pair (Fonollosa, 2019), or by leveraging asymmetries (Daniušis et al., 2010).

Despite the breadth of techniques available, it is important to note statistical dependence does

not imply causal dependence, i.e., the causal graphs that can be obtained might not be complete nor unique (Zhang et al., 2011; Weinberger, 2018). Thus, incorporating domain-specific or task-specific assumptions is fundamental to have satisfactory graphs (Hyvärinen and Pajunen, 1999; Zhang et al., 2015).

B Crowdsourcing Setup

Here, we outline our crowdsourcing setup for curating candidate SHOULD-KNOW obtained with scene graph generation (subsection 4.1). We recruited 520 workers (avg. wage: 8 GBP/h) on Prolific from English-speaking countries, with an approval rate of $\geq 95\%$. This study received ethics approval from our institution (ID: 4696).

Before carrying out the tasks, participants were required to provide explicit consent (which they could revoke at any time during the study without penalty) and to complete a brief tutorial explaining the task and the requirements.

B.1 Verification Task

For this task, participants check the correctness and the relevance of the triplets extracted in the SGG step with respect to the input image and text (Figure 6). Participants can confirm or edit concept labels, concept locations, and relationship labels. They are assisted in copy-editing the triplets with a lightweight auto-complete mechanism based on the concept and relationship labels used by IETrans. Note that participants were not limited to these labels and could provide new ones (Figure 7).

B.2 Expansion Task

In this task, we seek to obtain additional triplets that were not identified in the SGG step. We asked participants to indicate these relevant but missing triplets to extend the previously verified SK specifications, if needed (Figure 8). Here, human annotations comprise concept labels, bounding boxes, and relationship labels.

C Prompt Templates

In this section, we outline the prompt templates designed to obtain responses and concomitant self-explanations from VLMs (subsection 4.2). These result from several refinements using ChatGPT and manual checking to ensure their correctness, adherence to the task, and lack of hallucinations.

We provide the prompt templates we used for

Dataset	$J_{concepts}$	$J_{relations}$
LLaVa-Bench	0.81	0.82
MMBench – Image Scene	0.41	0.91
MMBench – Image Topic	0.99	0.90
SEED-Bench 2 – Scene Understanding	0.85	0.88
SEED-Bench 2 – Visual Reasoning	0.99	0.94
VQAv2 – How many people are...	0.47	0.89
VQAv2 – What is the person...	0.45	0.87

Table 4: Average Jaccard distance, for each dataset, measured between the sets of unique concepts and relationships obtained with and without USE_VISION enabled.

- Collecting model responses: Figure 9
- Producing rationales: Figure 10
- Structuring rationales: Figure 11

D Additional Implementation Details

D.1 BERT Score Configuration

We followed the indications of the Zhang et al. (2020) and used DeBERTa-XLarge to compute the BERT Score measures, given its better correlation with human evaluators. We report the hashcode to show the settings we used for BERT Score (Zhang et al., 2020).

```
microsoft/deberta-xlarge-mnli_L40
_no-idf_version=0.3.12(hug_trans=4.30.0)
```

D.2 Double ML Setup

For our DML in DiaVLo, we use the implementations from DoWhy and EconML. These implement the DML2 variant from (Chernozhukov et al., 2018). These packages handle sample splitting and cross-fitting, which are crucial in the DML framework. We use the default number of 2 sample splits. Finally, we perform 5-fold cross-validation to estimate the causal effects from the residuals. For the two regressors, we use gradient-boosted trees to estimate $\hat{f}(Z)$ and $\hat{g}(Z)$. We rely on the scikit-learn implementation GradientBoostingRegressor, with the following parameters:

- Number of estimators = 100
- Tree depth = 3
- Learning rate = 0.1

Overview

Progress: 20.0%

[Instructions for Validation Step](#) [Opt Out](#)

For viewing images and concepts in the best way possible, please do not resize this window.

Once you are done with the current image-question pair, click on the 'Submit Annotations' button below to proceed.

Question: What is funny about this image? Describe it panel by panel.

[Submit Annotations](#)



[TO DO]	Triple #1: <floor, has, line>	▼
[TO DO]	Triple #2: <screen, shows, screen>	▼
[TO DO]	Triple #3: <button, on, shirt>	▼
[TO DO]	Triple #4: <finger, holding, remote>	▼
[TO DO]	Triple #5: <hand, hand of, person>	▼
[TO DO]	Triple #6: <remote, held by, finger>	▼
[TO DO]	Triple #7: <hand, holding, remote>	▼
[TO DO]	Triple #8: <screen, on, screen>	▼
[TO DO]	Triple #9: <person, striped, shirt>	▼
[TO DO]	Triple #10: <finger, belonging to, person>	▼
[TO DO]	Triple #11: <remote, held by, person>	▼

(Make sure that all triples are marked as **[OK]** by scrolling through the list.)

Figure 6: Verification Task

Triple #1: (hand, hand of, person) **[OK]** ▼

Triple #2: (person, shaking, hand) **[TO DO]** ▲

Q1 - Is the triple useful to answer the Question?
☒ Yes ☐ No

Q2 - Are the entities in the triple properly identified in the picture?
☐ Yes ☒ No

Entity #1
 Label: [Change Bounding Box](#) (click and drag to draw)

Entity #2
 Label: [Change Bounding Box](#) (click and drag to draw)

Q3 - Is the relationship between the two entities properly defined?

Figure 7: Form to edit triple data with the auto-complete suggestions visible (highlighted).

E Additional Results

E.1 Ablating IETrans

We verified that IETrans appropriately relies on visual inputs rather than on spurious linguistic correlations stemming from transferring data from general predicate labels to specific ones (i.e., internal transfer) or relabelling (i.e., external transfer). See

the original work by (Zhang et al., 2022a) for the details on IETrans data transfer.

Concretely, we compared IETrans runs when PREDICT_USE_VISION is set to False (ablation) and True (our experiments). For this, we measure the difference in the number of unique concepts detected by IETrans (Table 5), as well as the Jaccard distance between the unique concept and relationship labels extracted in the two settings (Table 4). We see that once the visual signal is turned off, IETrans identifies fewer concepts and assigns very different labels to any given sample across the datasets considered.

E.2 Behaviour Threshold Sensitivity

In addition to Figure 3, we report in Figure 12 the proportions of samples classified as Aligned, Expanded, or Divergent as τ_a and τ_e vary.

E.3 Ablating OWLv2

We applied OWLv2 (Minderer et al., 2023) zero-shot with the default parametrisation to help us match the SK and RK specifications and derive bounding boxes for visual concepts in the self-explained RK. We considered the top-1 bounding box from OWLv2, using the suggested post-

Overview

Progress: 20.0%

[Instructions for Annotation Step](#) [Opt Out](#)

For viewing images and concepts in the best way possible, please do not resize this window.

Click on the 'New Annotation' button to provide additional new annotations for the question-image pair.

Once you are done with the current image-question pair, click on the 'Submit Annotations' button to proceed.

Question:
What is funny about this image? Describe it panel by panel.

[Submit Annotations](#) [New Annotation](#)



[TO DO] New annotation #1:

New Entity #1
Label:
[Set Bounding Box](#) (click and drag over the image to draw)

Predicate
New Predicate:

New Entity #2
Label:
[Set Bounding Box](#)

[TO DO] New annotation #2:

[TO DO] New annotation #3:

(Make sure that all triples are marked as **[OK]** by scrolling through the list.)

Figure 8: Verification Task

processing threshold $\tau_{\text{OWLv2}} = 0.1$ and suppression threshold $\tau_{\text{nms}} = 0.3$ used to collate overlapping bounding boxes. Table 6 shows the intermediate percentages of concepts for which bounding boxes were found throughout that step of DiaVLo. To understand how OWLv2 behaves within DiaVLo, we carried out two ablations:

- We swept τ_{OWLv2} within its range $[0.0, 1.0]$. As we increase τ_{OWLv2} , we can see two phenomena (see Figure 13). First, higher τ_{OWLv2} values cause more and more REALLY-KNOW triplets to be erroneously passed onto the semantic fallback step, i.e., the third and last step of our SK-RK matching setup. This is expected as we match SK and RK incrementally. Second, the SK-RK may decrease to a point where OWLv2 matches fewer triplets compared to exact, lexicon-based matching (e.g., for LLaVa-1.6 on MMBench), therefore missing out on relevant triplets. This may be due to poor calibration of OWLv2’s own scores.
- We swept the internal Non Maximum Suppression (NMS) threshold τ_{nms} , while fixing $\tau_{\text{OWLv2}} = 0.1$. The results (Figure 14) show that, for the most part, the overall number of

SK-RK matches starts to plateau for $\tau_{\text{nms}} > 0.3$. High τ_{nms} may also lead to bounding boxes being aggregated too coarsely, thereby losing information. These results confirm that, in our specific case, the default $\tau_{\text{nms}} = 0.3$ is an effective sweet spot.

E.4 Mutual Information Bootstrapping

Given our small sample size, we use bootstrapping to examine MI itself and assess potential biases and overestimates. Concretely, we bootstrap p-values and 95% confidence intervals and perform permutation checks. We perform 1000 resampling draws and correct our MI estimates as

$$\text{MI}_{\text{corrected}} = \text{MI}_{\text{observed}} - \text{MI}_{\text{permutation}} \quad (7)$$

We report the results in Table 7.

Another important point is that we use the scikit-learn implementation of MI (mutual_information_regression and mutual_information_classif), which is based on entropy estimation using k-nearest-neighbour distances. We visually verify the MI estimates for $k = [3, 5, 7, 10, 15, 20, 30]$ to determine a range suitable for robust MI estimation. From the MI

Collecting model responses
LLaVa-Bench (same for all models) {{ image and question from dataset }}
MMBench (same for all models) Provide a concise and descriptive caption for this image. {{ image and question from dataset }}
SEED-Bench 2 (for InternVL2, LLaVa-1.6, and Qwen2.5-VL) Based on the provided image, analyze the given question and choose the most accurate option from the listed alternatives. Your response should consist solely of the letter corresponding to the correct answer (A, B, C, or D), without any additional explanation. {{ image and question from dataset }}
SEED-Bench 2 (for ShareGPT4V) Based on the provided image, analyze the given question and choose the most accurate option from the listed alternatives. Your response <u>must</u> consist solely of <u>the answer you believe is correct</u> , without any additional explanation. {{ image and question from dataset }}
VQA v2 (same for all models) Based on the given image, answer the question using only one word. Provide no additional text or explanation. {{ image and question from dataset }}

Figure 9: Prompt templates for getting model responses. Model-specific changes are underlined.

estimates shown in [Figure 15](#), [Figure 16](#), [Figure 17](#), and [Figure 18](#), we can determine that:

- Given that MI estimates collapse to ≈ 0 when we randomly shuffle the data, measured performance is indeed related to the behaviours identified; and,
- For $k \in [5, 10]$, we have a practical and balanced range that is (1) less sensitive to noise and overestimations (low k) and (2) less aggressively smoothed (high k).

To provide robust MI estimates, we take the median MI for $k \in \{5, 6, 7, 8, 9, 10\}$.

E.5 Examples of Model Behaviours

- Aligned behaviours: [Figure 19](#), [Figure 20](#)
- Expanded behaviours: [Figure 21](#), [Figure 22](#), [Figure 23](#), [Figure 24](#)
- Divergent behaviours: [Figure 25](#), [Figure 26](#)

E.6 Model-pairwise Distributions of Causal Effects

- Distributions on LLaVa-Bench: [Figure 27](#)

- Distributions on MMBench: [Figure 28](#)
- Distributions on SEED-Bench-2: [Figure 29](#)
- Distributions on VQA v2: [Figure 30](#)

Producing rationales

Review the given answer and carefully analyze the image. Identify and describe the key visual elements and concepts in the image that directly support your reasoning and answer. For each point, clearly explain how that visual element contributes to the answer, and ensure every point is directly linked to the reasoning behind the answer.

Present your rationale in a structured, bullet-point format with the following guidelines:

- Each bullet point must begin with the * character.
- Each point should clearly explain the connection between a specific visual element and the given answer.
- Avoid general descriptions of the image; focus on visual details that are relevant to the answer.
- Keep each point concise and ensure it directly addresses the reasoning behind the answer.

Figure 10: Prompt templates for generating model rationales. Same for all models.

Dataset	Concept Detected		Relations Detected	
	w/ USE_VISION	w/o USE_VISION	w/ USE_VISION	w/o USE_VISION
LLaVa-Bench	746	689 (↓ 7.64%)	720	720 (-)
MMBench – <i>Image Scene</i>	7905	6393 (↓ 19.13%)	7650	7650 (-)
MMBench – <i>Image Topic</i>	1980	179 (↓ 90.96%)	1920	1920 (-)
SEED-Bench 2 – <i>Scene Understanding</i>	100903	71883 (↓ 28.76%)	94740	94740 (-)
SEED-Bench 2 – <i>Visual Reasoning</i>	10715	1314 (↓ 87.74%)	9930	9930 (-)
VQAv2 – <i>How many people are...</i>	58894	40604 (↓ 31.06%)	55140	55110 (↓ 0.05%)
VQAv2 – <i>What is the person...</i>	26136	19292 (↓ 26,19%)	25530	25530 (-)

Table 5: IETrans concepts and relationships extracted when the vision signal is used (w/ USE_VISION) compared to when it is not (w/o USE_VISION).

VLM	Dataset	Initial RK	Exact Matching	OWLv2 Matches	Semantic Fallback	Matched	Not Matched
InternVL2	LLaVa-Bench	605	0	385	220	605	0
	MMBench	1341	3	1000	338	1341	0
	SEED-Bench 2	1388	3	1030	355	1388	0
	VQA v2	1186	15	1009	162	1186	0
LLaVa-1.6	LLaVa-Bench	421	0	235	186	421	0
	MMBench	1255	39	828	388	1255	0
	SEED-Bench 2	1192	2	966	224	1192	0
	VQA v2	1364	23	1146	195	1364	0
Qwen2.5-VL	LLaVa-Bench	546	0	159	387	546	0
	MMBench	620	12	476	132	620	0
	SEED-Bench 2	564	0	248	316	564	0
	VQA v2	346	4	211	131	346	0
ShareGPT4V	LLaVa-Bench	423	0	260	163	423	0
	MMBench	928	12	651	265	928	0
	SEED-Bench 2	311	1	265	35	301	10
	VQA v2	58	2	49	7	58	0
Total Matched		12548	116	8918	3504	12538	10
% Matched			0.92%	71.07%	27.29%	99.92%	0.08%

Table 6: Intermediate RK-SK matching counts and percentages.

Structuring rationales

InternVL2 and Qwen2.5-VL

Carefully analyze each point in the list above and extract distinct pairs of entities and their relationships. For each point, identify exactly two distinct entities at a time, along with a detailed and context-specific relationship between them (e.g., action, spatial position, interaction, causality, part-whole relationship, etc.). For each extraction, you must use the following format exactly: (Entity: [entity_1], Relationship: [relationship], Entity: [entity_2]).

Additional guidelines:

- Both entities must always be present: Ensure that each extraction contains two distinct entities in the format provided.
- Do not include references to the image itself (e.g., 'image,' 'picture,' 'photo') as an entity.
- Avoid overly simplistic relationships (e.g., 'located in' or 'visible in'). Focus on specific interactions or connections, such as 'forms,' 'comprises,' 'is part of,' 'affects,' or 'is near.'
- Focus on distinct entity pairs, avoiding duplicates across all points.
- If a point contains multiple entities or relationships, break them down into separate pairs.
- Entities can refer to physical objects, people, animals, abstract concepts, etc.
- Relationships should reflect more complex interactions or spatial/temporal arrangements.
- Ensure the relationship clearly reflects how the entities are connected in context. Consider relationships like causality, containment, or part-whole structure.
- Prioritize clarity and contextual relevance in your extractions.
- Each extraction must contain two distinct entities; avoid incomplete extractions with only one entity.

Example:

- Original sentence: 'A dog is playing with a ball near the tree.'
- Output:
 - (Entity: dog, Relationship: playing with, Entity: ball)
 - (Entity: dog, Relationship: near, Entity: tree)
 - (Entity: tree, Relationship: part of, Entity: landscape)

LLaVa-1.6 and ShareGPT4V

Analyze the provided content and extract distinct pairs of entities and their relationships using the guidelines below:

1. Output Format: Each extraction must follow this format: * (Entity: [entity_1], Relationship: [relationship], Entity: [entity_2])

2. Entity Guidelines:

- Extract two distinct entities for each pair.
- Exclude references to the image itself (e.g., "image," "picture").
- Entities can be physical objects, people, animals, or abstract concepts.

3. Relationship Guidelines:

- Identify meaningful, context-specific relationships (e.g., forms, comprises, is part of, affects, is near).
- Relationships should describe interactions, connections, or arrangements between the entities, such as causality, spatial/temporal proximity, or part-whole structures.

4. Rules for Clarity and Consistency:

- Each extraction must include exactly two entities and one relationship. Avoid incomplete or ambiguous pairs.
- Break down complex points into multiple pairs if needed.
- Avoid duplicates across extractions.

5. Relevance and Accuracy:

- Ensure extractions are contextually accurate and relevant.
- Prioritize clear, precise descriptions of entity relationships.

Example:

- Original Sentence: "A dog is playing with a ball near the tree."
- Extractions:
 - (Entity: dog, Relationship: playing with, Entity: ball)
 - (Entity: dog, Relationship: near, Entity: tree)

Figure 11: Prompt templates for structuring VLM rationales. Changes are mostly related to how instructions are organised and their clarity.

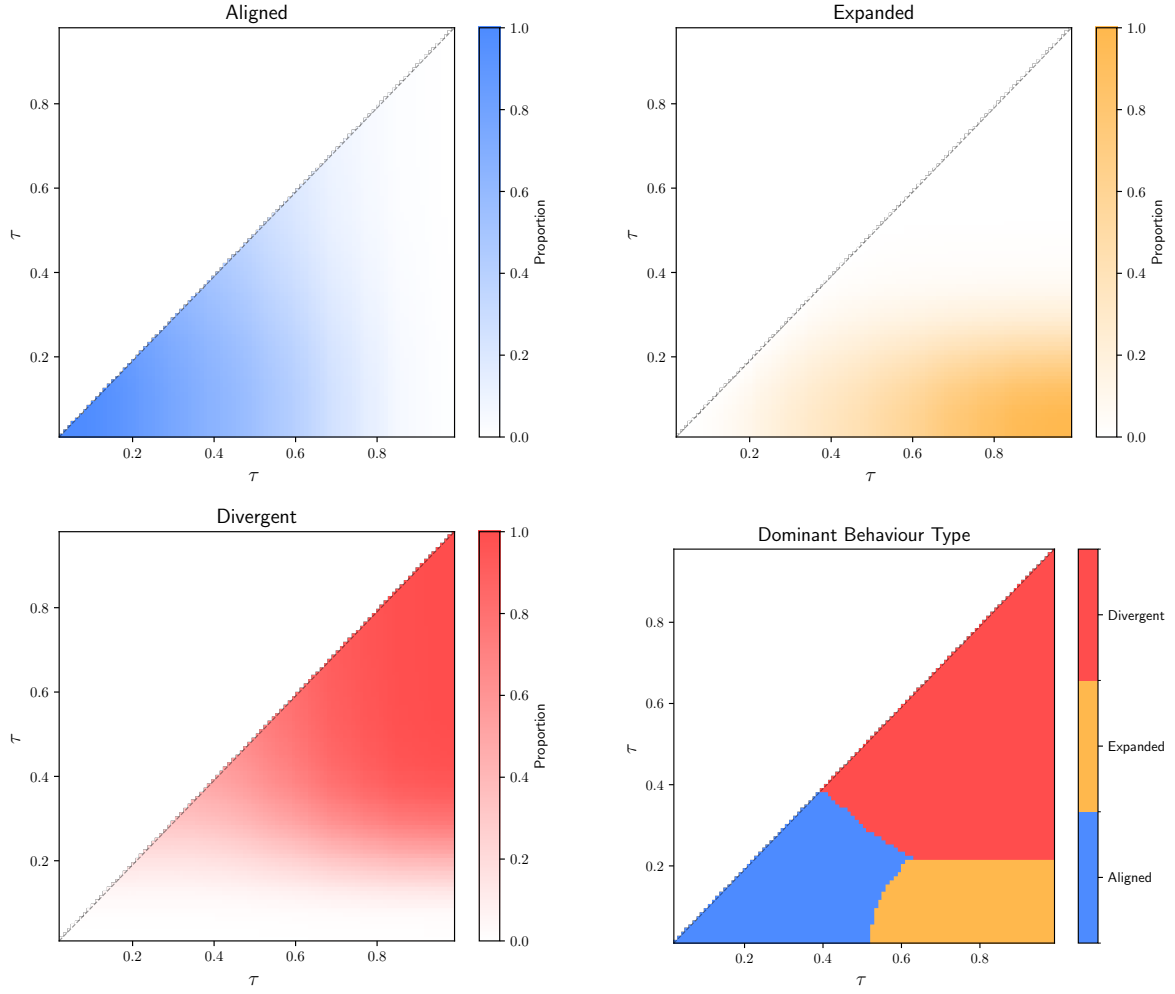


Figure 12: Proportions of samples classified as Aligned, Expanded, or Divergent, and dominant behaviour type as τ_a and τ_e vary.

Model	LLaVa-Bench		MMBench		SEED-Bench 2		VQA v2	
	A	MI	A	MI	A	MI	A	MI
InternVL2	0.673	3.367 [3.158, 3.589]	0.796	3.200 [3.053, 3.350]	0.787	0.379 [0.325, 0.437]	0.667	0.476 [0.433, 0.526]
LLaVa-1.6	0.657	2.389 [2.052, 2.711]	0.792	2.512 [2.333, 2.699]	0.667	0.438 [0.385, 0.497]	0.585	0.481 [0.437, 0.531]
Qwen2.5-VL	0.670	3.008 [2.764, 3.270]	0.743	1.561 [1.331, 1.792]	0.833	0.188 [0.115, 0.275]	0.700	0.147 [0.082, 0.230]
ShareGPT4V	0.687	2.058 [1.737, 2.405]	0.705	2.116 [1.932, 2.318]	0.661	0.292 [0.198, 0.404]	0.810	0.558 [0.139, 1.074]

Table 7: VLMs accuracy (A) on the four datasets and Mutual Information (MI) measures between accuracy and model behaviour types. In parentheses, we report the bootstrapped 95% confidence intervals for the MI measures.

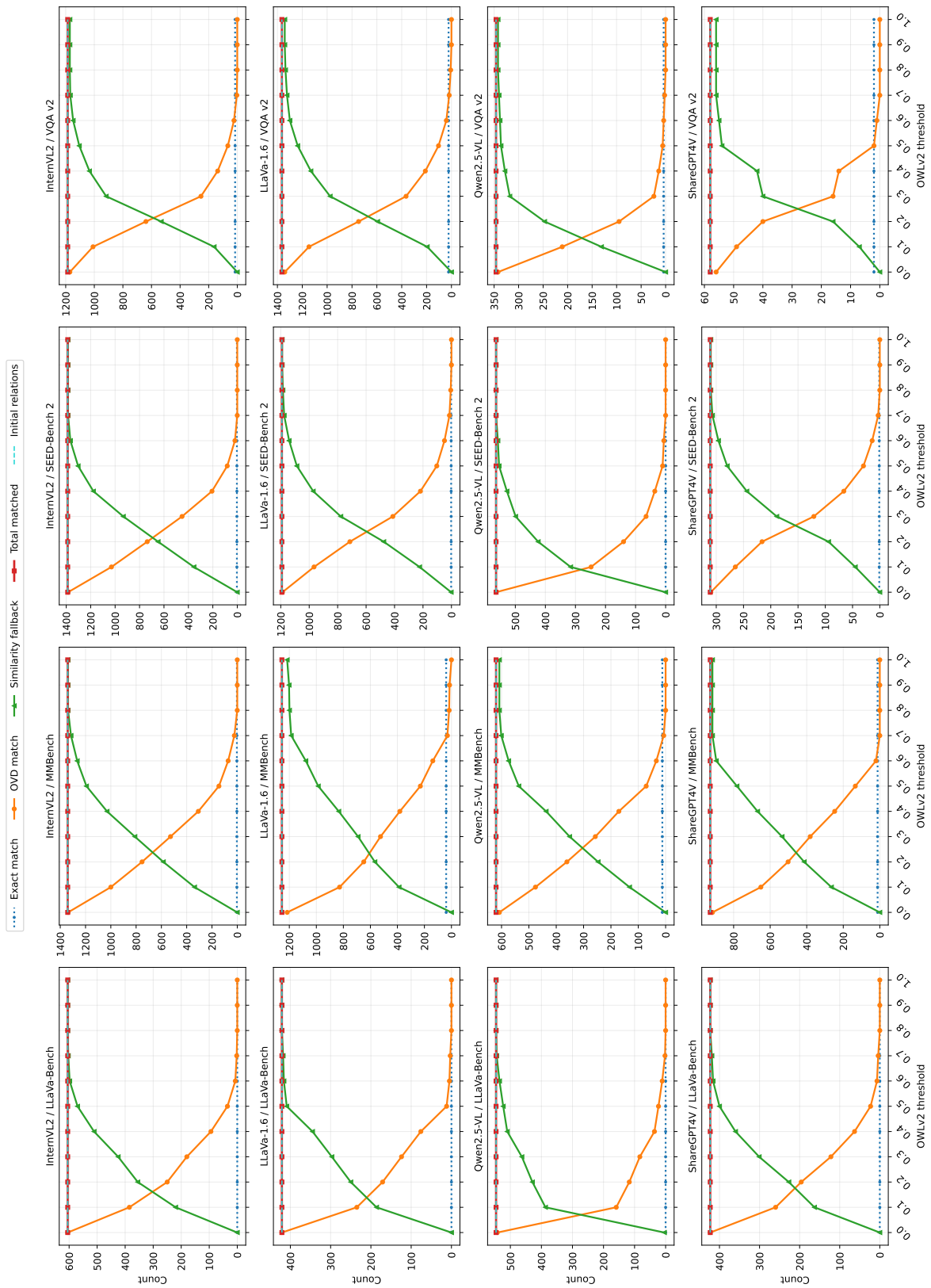


Figure 13: Results from τ_{OWLv2} sweep.

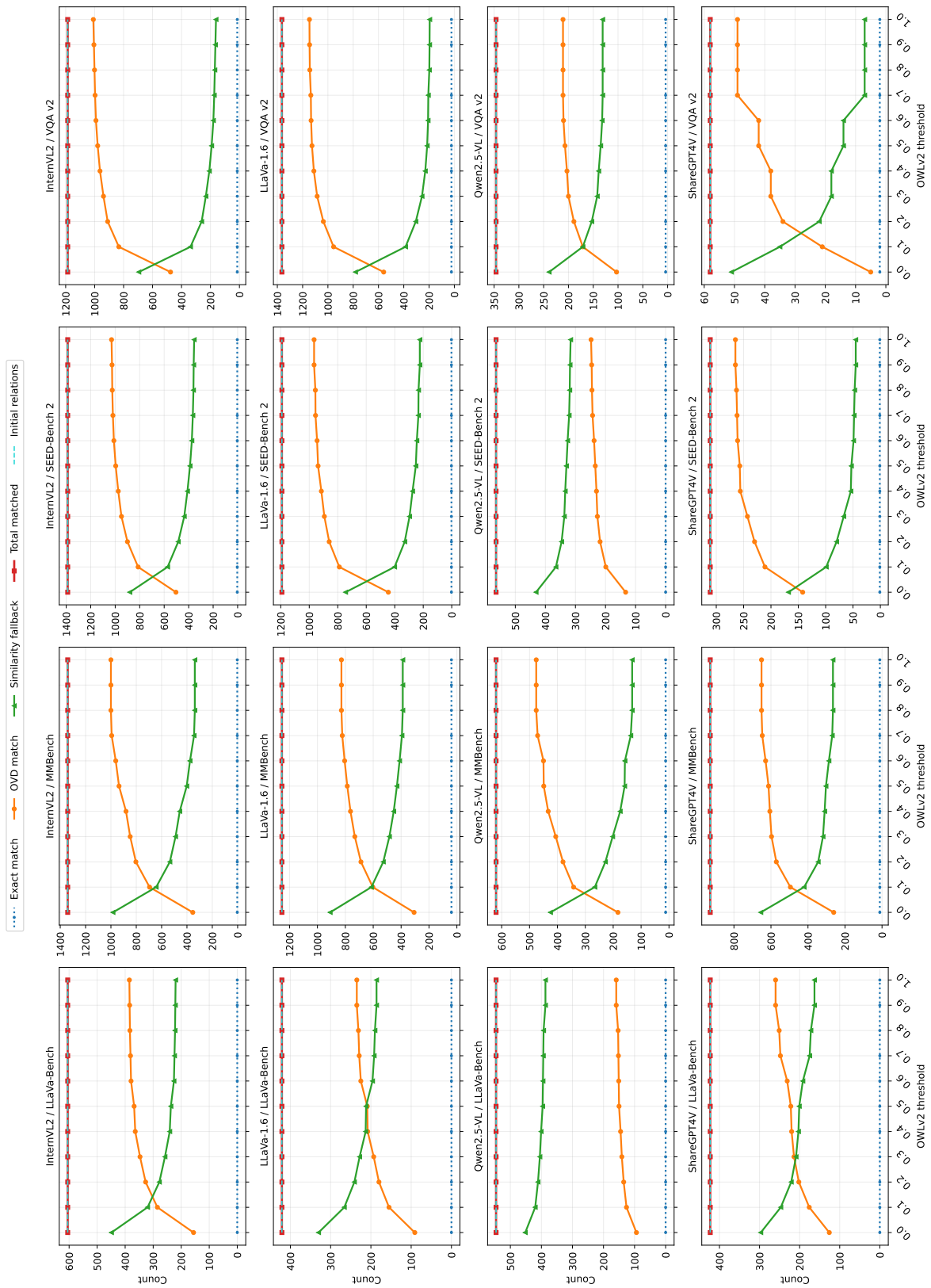


Figure 14: Results from τ_{rms} sweep.

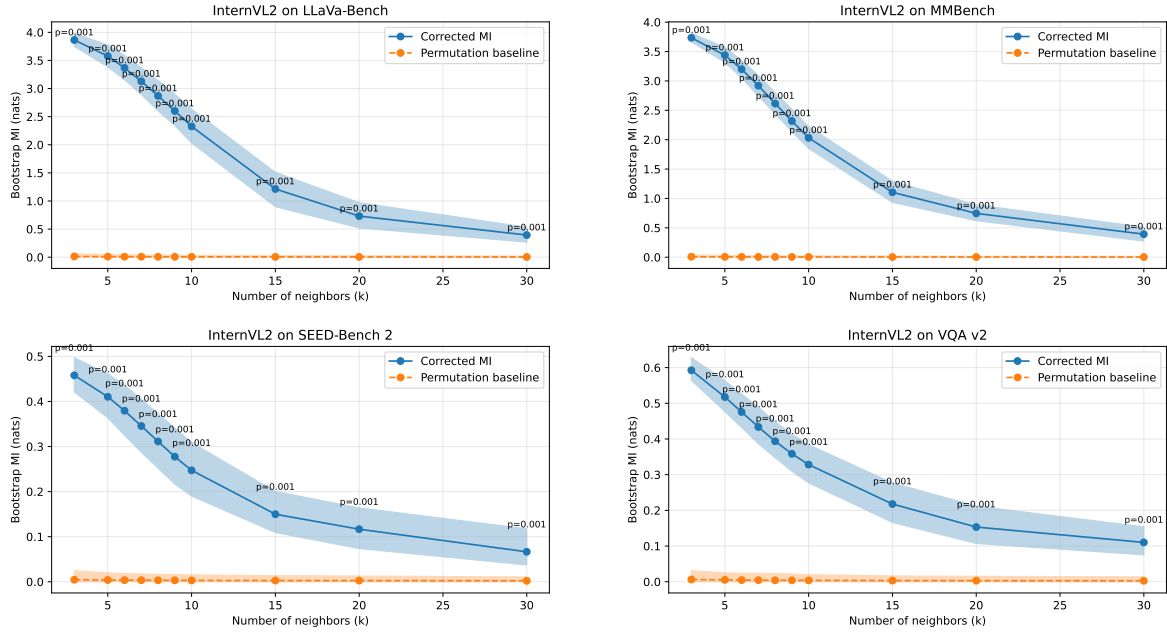


Figure 15: Bootstrapped Mutual Information measures InternVL2 (Chen et al., 2024).

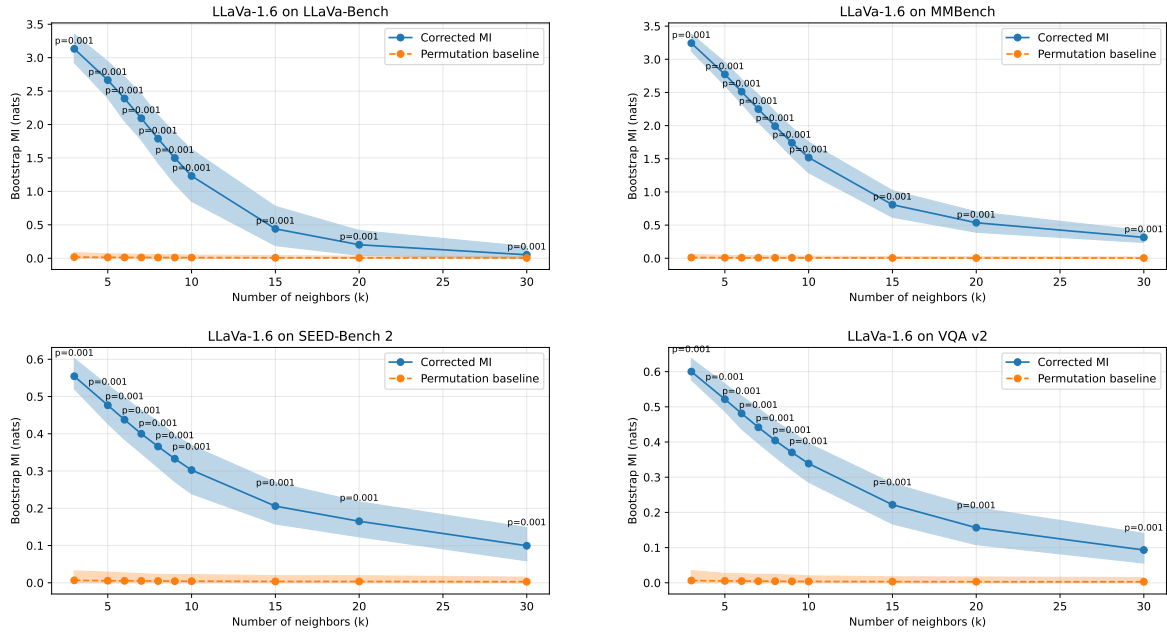


Figure 16: Bootstrapped Mutual Information measures LLaVa-1.6 (Liu et al., 2024b).

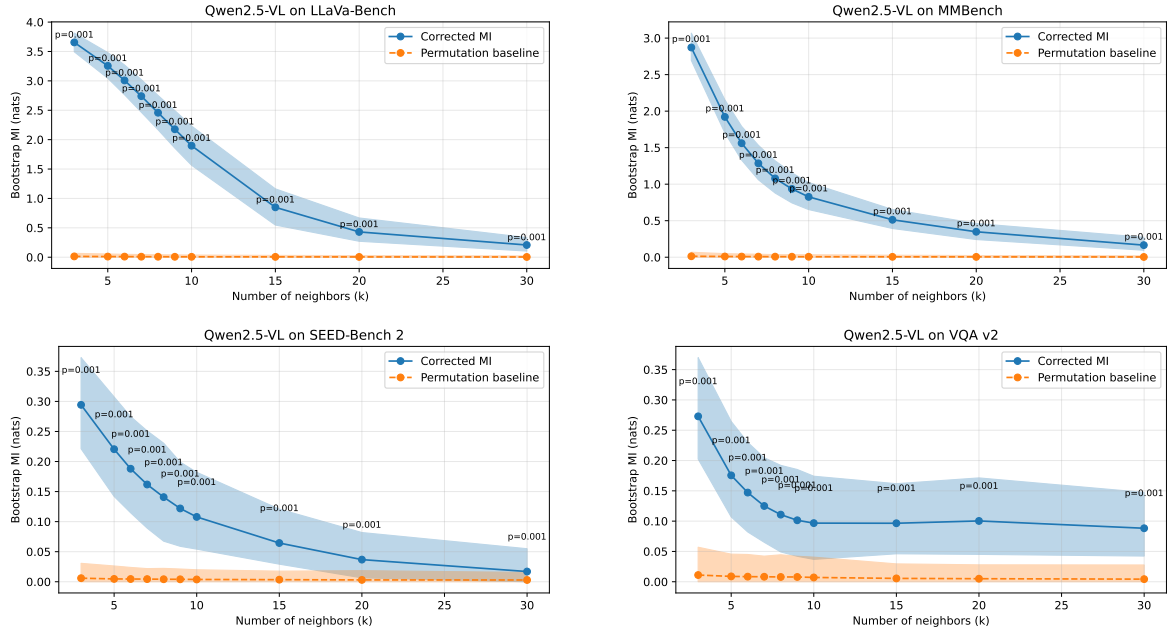


Figure 17: Bootstrapped Mutual Information measures Qwen2.5-VL (Bai et al., 2025).

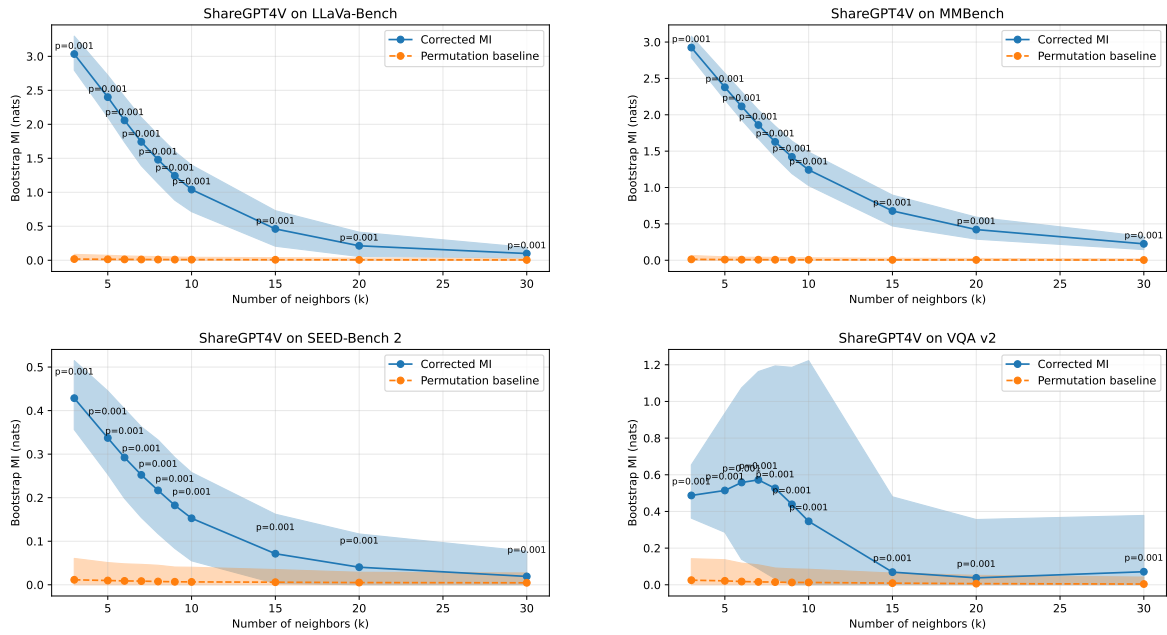


Figure 18: Bootstrapped Mutual Information measures ShareGPT4V (Chen et al., 2023).

LLaVa-1.6 on MMBench (Sample ID: 2000067)

Input image



REALLY-KNOWs

1. {"from_concept": "cat", "relationship": "laying in", "to_concept": "suitcase"}
 - a. SK match (0.964): {"from_concept": "cat", "relationship": "laying on", "to_concept": "suitcase"}
2. {"from_concept": "cat", "relationship": "has white and brown fur", "to_concept": "fur"}
 - a. SK match (0.557): {"from_concept": "cat", "relationship": "growing", "to_concept": "whisker"}

Input question

Provide a concise and descriptive caption for the given image.

Figure 19: Example of aligned model behaviour from LLaVa-1.6 on MMBench.

LLaVa-1.6 on VQA v2 (Sample ID: 470070003)

Input image



REALLY-KNOWs

1. {"from_concept": "umpire", "relationship": "wearing", "to_concept": "shirt"}
 - a. SK match (1.00): {"from_concept": "umpire", "relationship": "wearing", "to_concept": "shirt"}
2. {"from_concept": "umpire", "relationship": "wearing", "to_concept": "cap"}
 - a. SK match (0.924): {"from_concept": "umpire", "relationship": "wearing", "to_concept": "hat"}
3. {"from_concept": "umpire", "relationship": "wearing", "to_concept": "shoes"}
 - a. SK match (0.882): {"from_concept": "umpire", "relationship": "wearing", "to_concept": "white shoes"}
4. {"from_concept": "umpire", "relationship": "wearing", "to_concept": "uniform"}
 - a. SK match (0.801): {"from_concept": "umpire", "relationship": "wearing", "to_concept": "pant"}
5. {"from_concept": "umpire", "relationship": "standing", "to_concept": "field"}
 - a. SK match (0.698): {"from_concept": "umpire", "relationship": "has", "to_concept": "hand"}
6. {"from_concept": "umpire", "relationship": "wearing", "to_concept": "belt"}
 - a. SK match (0.685): {"from_concept": "umpire", "relationship": "wearing", "to_concept": "black hat"}

Input question

What is the person in the black shirt doing?

Figure 20: Example of aligned model behaviour from LLaVa-1.6 on VQA v2.

Qwen2.5-VL on LLaVa-Bench (Sample ID: 39)

Input image



REALLY-KNOWs

1. {"from_concept": "adapter", "relationship": "fits onto", "to_concept": "lightning cable"}
 - a. SK match (0.735): {"from_concept": "lightning plug", "relationship": "on", "to_concept": "connector"}
2. {"from_concept": "adapter", "relationship": "shown in", "to_concept": "packaging"}
 - a. SK match (0.465): {"from_concept": "package", "relationship": "containing", "to_concept": "blue vga connector"}
3. {"from_concept": "cable", "relationship": "has adapter", "to_concept": "adapter"}
 - a. SK match (0.459): {"from_concept": "white cable", "relationship": "connected to", "to_concept": "blue vga connector"}
4. {"from_concept": "phone", "relationship": "connected to", "to_concept": "cable"}
 - a. SK match (0.416): {"from_concept": "connector's cable", "relationship": "on", "to_concept": "background cloth"}
5. {"from_concept": "adapter", "relationship": "has latch mechanism", "to_concept": "latch mechanism"}
 - a. SK match (0.379): {"from_concept": "connector", "relationship": "held by", "to_concept": "finger"}

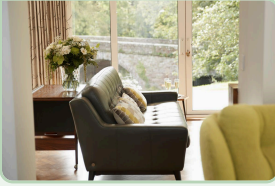
Input question

Give a detailed description of this image. Describe it panel by panel.

Figure 21: Example of expanded model behaviour from Qwen2.5-VL on LLaVa-Bench.

LLaVa-1.6 on SEED-Bench-2 (Sample ID: 4039)

Input image



Input question

What is the most likely reason for the glass door in the living room to be open?

REALLY-KNOWs

1. {"from_concept": "flowers", "relationship": "is in", "to_concept": "vase"}
 - a. SK match (0.860): {"from_concept": "flower", "relationship": "blooming in", "to_concept": "vase"}
2. {"from_concept": "vase", "relationship": "is on", "to_concept": "table"}
 - a. SK match (0.677): {"from_concept": "flowers", "relationship": "on", "to_concept": "table"}
3. {"from_concept": "door", "relationship": "is part of", "to_concept": "living room"}
 - a. SK match (0.512): {"from_concept": "tree", "relationship": "outside", "to_concept": "door"}
4. {"from_concept": "balcony", "relationship": "is near", "to_concept": "door"}
 - a. SK match (0.433): {"from_concept": "fresh air", "relationship": "outside", "to_concept": "window"}
5. {"from_concept": "potted plant", "relationship": "is on", "to_concept": "table"}
 - a. SK match (0.434): {"from_concept": "plant", "relationship": "outside", "to_concept": "window"}
6. {"from_concept": "balcony", "relationship": "is part of", "to_concept": "living room"}
 - a. SK match (0.327): {"from_concept": "tree", "relationship": "outside", "to_concept": "window"}
7. {"from_concept": "table", "relationship": "is near", "to_concept": "couch"}
 - a. SK match (0.129): {"from_concept": "sunlight", "relationship": "outside", "to_concept": "window"}
8. {"from_concept": "couch", "relationship": "is part of", "to_concept": "living room"}
 - a. SK match (-): -

Figure 22: Example of expanded model behaviour from LLaVa-1.6 on SEED-Bench-2.

LLaVa-1.6 on MMBench (Sample ID: 3001808)

Input image



Input question

Provide a concise and descriptive caption for the given image.

REALLY-KNOWs

1. {"from_concept": "pedicure tool", "relationship": "on", "to_concept": "foot"}
 - a. SK match (0.546): {"from_concept": "woman", "relationship": "getting", "to_concept": "foot massage"}
2. {"from_concept": "women", "relationship": "providing", "to_concept": "pedicure"}
 - a. SK match (0.535): {"from_concept": "women", "relationship": "in", "to_concept": "spa"}
3. {"from_concept": "spa or salon", "relationship": "setting", "to_concept": "pedicure"}
 - a. SK match (0.506): {"from_concept": "spa worker", "relationship": "giving a massage", "to_concept": "client"}
4. {"from_concept": "chair", "relationship": "in", "to_concept": "spa or salon"}
 - a. SK match (0.309): {"from_concept": "leg", "relationship": "against", "to_concept": "wall"}
5. {"from_concept": "uniforms", "relationship": "matching", "to_concept": "women"}
 - a. SK match (0.023): {"from_concept": "head", "relationship": "resting on", "to_concept": "arm"}

Figure 23: Example of expanded model behaviour from LLaVa-1.6 on MMBench.

InternVL2 on VQAv2 (Sample ID: 2613006)

Input image



Input question

How many people are swimming?

REALLY-KNOWs

1. {"from_concept": "person", "relationship": "near", "to_concept": "wave"}
 - a. SK match (0.721): {"from_concept": "man", "relationship": "going for", "to_concept": "wave"}
2. {"from_concept": "surfboard", "relationship": "part of", "to_concept": "person"}
 - a. SK match (0.629): {"from_concept": "man", "relationship": "riding", "to_concept": "surfboard"}
3. {"from_concept": "person", "relationship": "swimming", "to_concept": "ocean"}
 - a. SK match (0.450): {"from_concept": "man", "relationship": "dressed", "to_concept": "wetsuit"}
4. {"from_concept": "ocean", "relationship": "part of", "to_concept": "wave"}
 - a. SK match (0.445): {"from_concept": "ripple", "relationship": "flowing in", "to_concept": "water"}

Figure 24: Example of expanded model behaviour from InternVL2 on VQA v2.

InternVL2 on VQA v2 (Sample ID: 63939005)

Input image



Input question

How many people are probably sitting at this table?

REALLY-KNOWs

```

1. {"from_concept": "table", "relationship": "occupied by", "to_concept": "people"}
  a. SK match (0.872): {"from_concept": "people", "relationship": "around", "to_concept": "table"}
2. {"from_concept": "table", "relationship": "set for", "to_concept": "eight people"}
  a. SK match (0.711): {"from_concept": "six people", "relationship": "at", "to_concept": "table"}
3. {"from_concept": "table", "relationship": "comprises", "to_concept": "place settings"}
  a. SK match (-): -
4. {"from_concept": "place settings", "relationship": "includes", "to_concept": "plates"}
  a. SK match (-): -
5. {"from_concept": "place settings", "relationship": "includes", "to_concept": "glasses"}
  a. SK match (-): -
6. {"from_concept": "place settings", "relationship": "includes", "to_concept": "cutlery"}
  a. SK match (-): -

```

Figure 25: Example of divergent model behaviour from InternVL2 on VQA v2.

ShareGPT4V on LLaVa-Bench (Sample ID: 33)

Input image



Input question

Write an attractive product description for this.

REALLY-KNOWs

```

1. {"from_concept": "mug", "relationship": "white", "to_concept": "ceramic"}
  a. SK match (0.559): {"from_concept": "letter", "relationship": "written on", "to_concept": "mug"}
2. {"from_concept": "mug", "relationship": "arranged", "to_concept": "other mugs"}
  a. SK match (0.497): {"from_concept": "cup", "relationship": "contains", "to_concept": "picture"}
3. {"from_concept": "mug", "relationship": "slightly tilted", "to_concept": "left"}
  a. SK match (0.241): {"from_concept": "cap", "relationship": "covering", "to_concept": "head"}
4. {"from_concept": "mug", "relationship": "detailed illustration", "to_concept": "mario"}
  a. SK match (-): -
5. {"from_concept": "mug", "relationship": "color", "to_concept": "vibrant"}
  a. SK match (-): -
6. {"from_concept": "mug", "relationship": "background", "to_concept": "dark blue"}
  a. SK match (-): -
7. {"from_concept": "mug", "relationship": "color", "to_concept": "contrast"}
  a. SK match (-): -
8. {"from_concept": "mug", "relationship": "arrangement", "to_concept": "composition"}
  a. SK match (-): -
9. {"from_concept": "mug", "relationship": "image", "to_concept": "background"}
  a. SK match (-): -
10. {"from_concept": "mug", "relationship": "focus", "to_concept": "illustration"}
  a. SK match (-): -

```

Figure 26: Example of divergent model behaviour from ShareGPT4V on LLaVa-Bench.

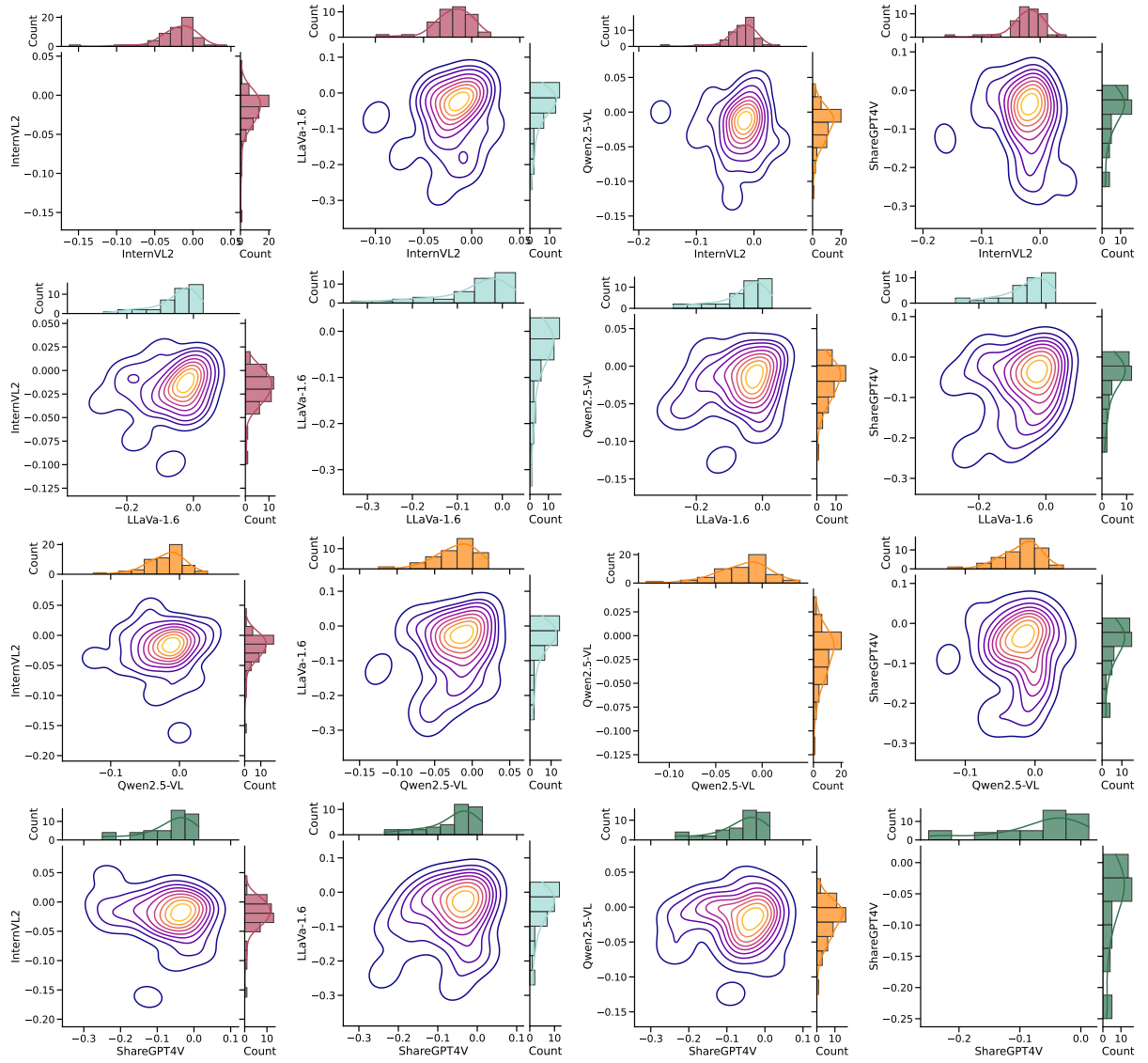


Figure 27: Model-pairwise distributions of estimated causal effects on LLaVa-Bench (Liu et al., 2023a).

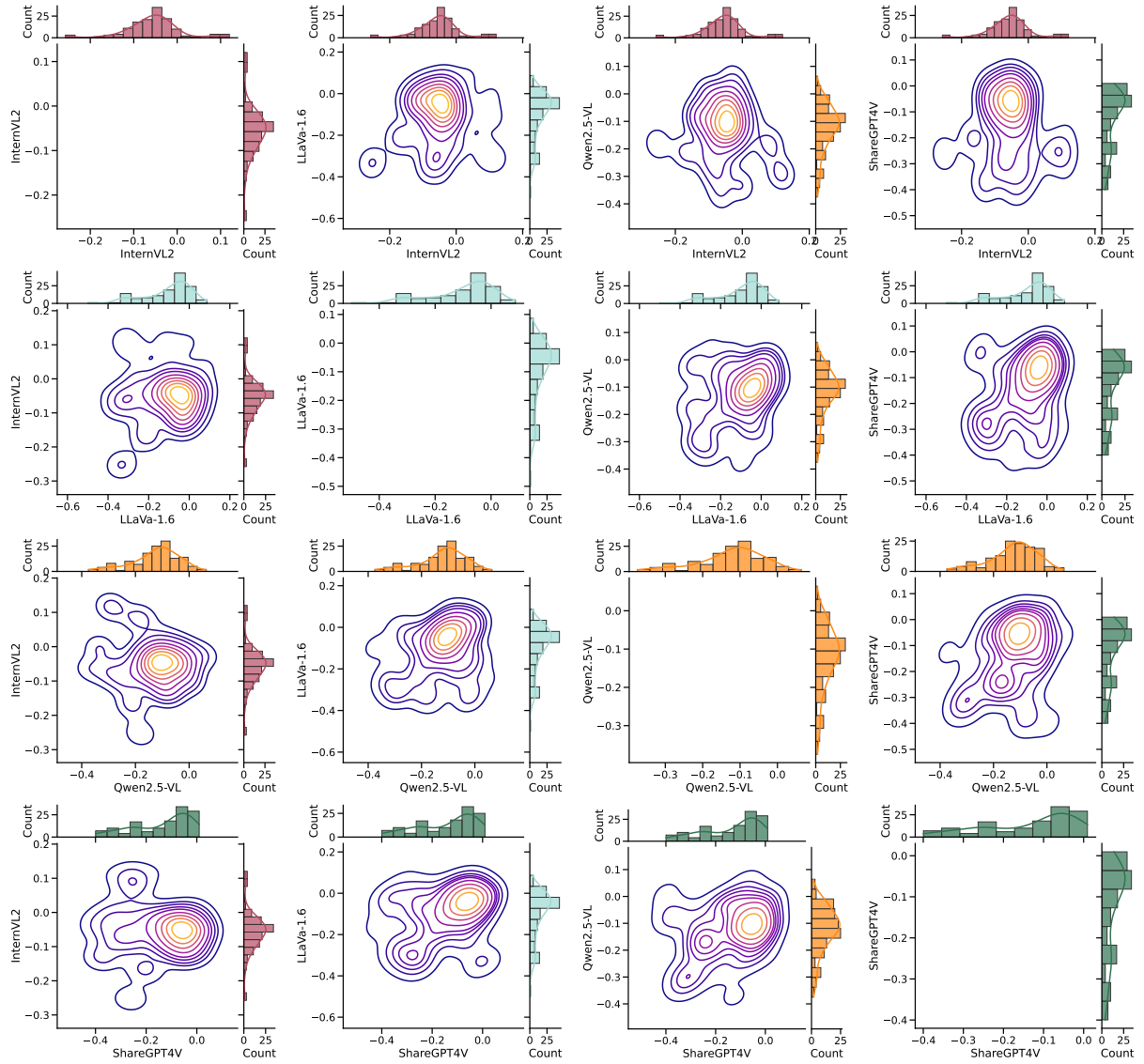


Figure 28: Model-pairwise distributions of estimated causal effects on MMBench (Liu et al., 2024c).

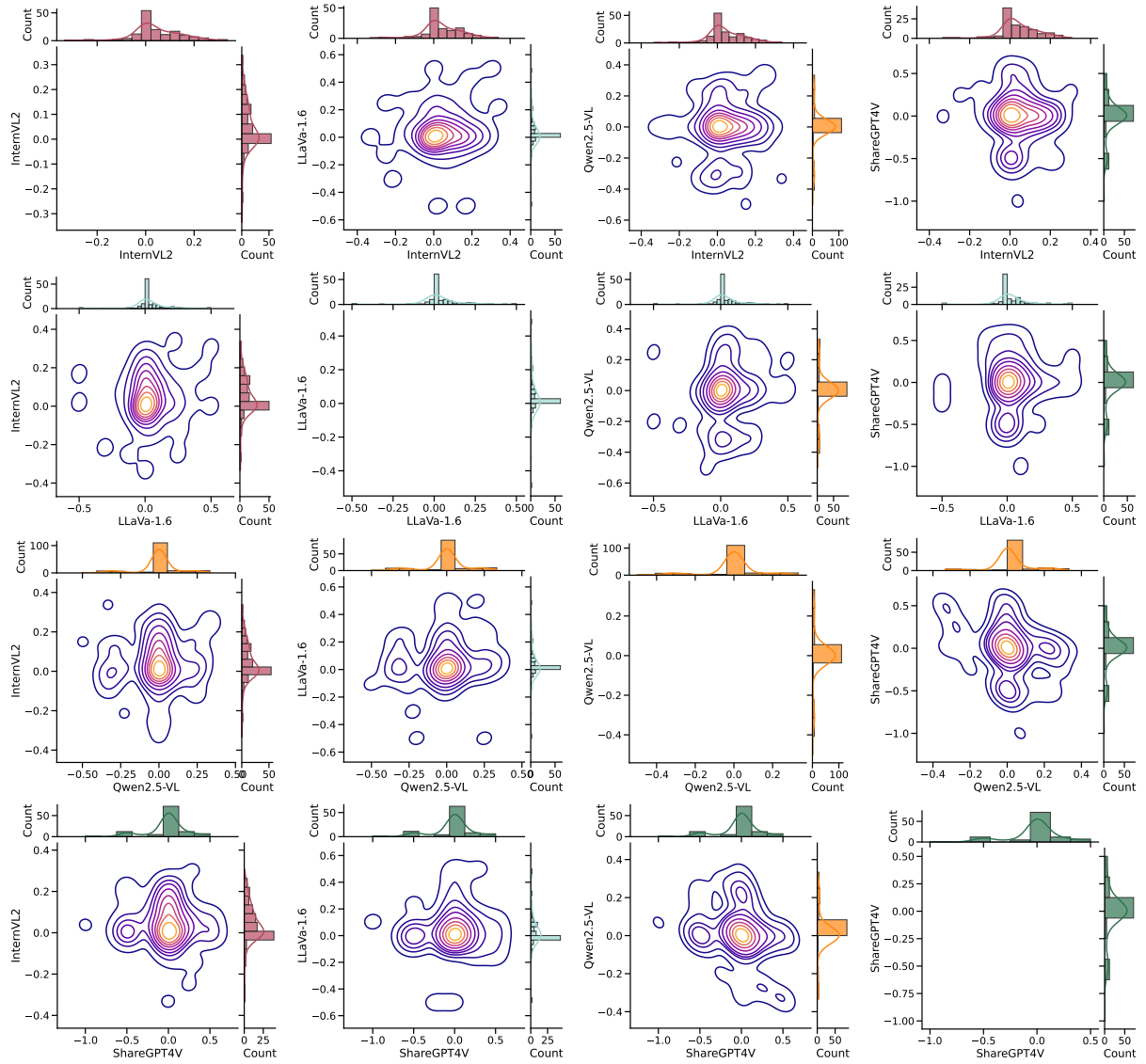


Figure 29: Model-pairwise distributions of estimated causal effects on SEED-Bench-2 (Li et al., 2024a).

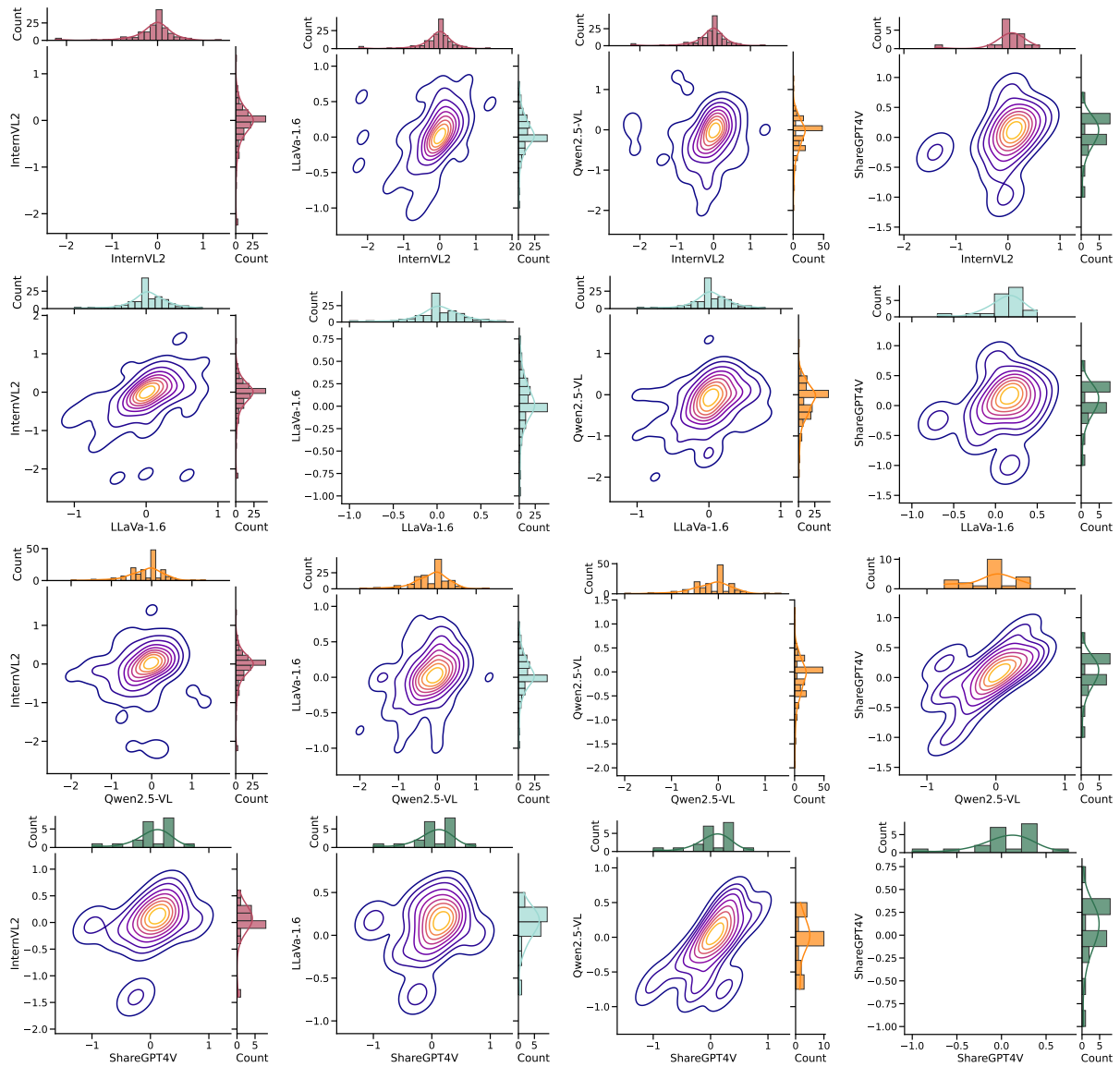


Figure 30: Model-pairwise distributions of estimated causal effects on VQA v2 (Goyal et al., 2017).