

Are Multimodal Benchmarks Really Useful? Item-Level Multimodal Benchmark Diagnosis via Structure-Response Co-Calibration

Shiqi Zhang¹ Weixin Zeng^{2*} Ziheng Zhang² Jiuyang Tang² Xiang Zhao²

¹National Key Laboratory of Information Systems Engineering,
National University of Defense Technology, Changsha, China

²National Key Laboratory of Big Data and Decision,
National University of Defense Technology, Changsha, China

{zhangshiqi, zengweixin13, zihengzhang20, jiuyang_tang, xiangzhao}@nudt.edu.cn

Abstract

Multimodal benchmarks are central to evaluating whether models can reason over information from multiple modalities. However, existing literature has revealed that some benchmarks may contain shortcut items that can be answered from a single modality or from answer-option regularities alone, thus failing to reflect models' true cross-modal inference ability. Existing methods for assessing benchmarks typically rely either on *intrinsic item structure* or *extrinsic model response behavior*, which are not adequate for a comprehensive diagnosis. To fill in this gap, we propose SRCoD, a structure-response co-calibrated diagnosis framework for item-level diagnosis of modality dependence in multimodal benchmarks. SRCoD estimates sample-wise multimodal information structure from benchmark content and target answers, and uses this signal as a soft structural anchor for modality-conditioned item response modeling. The learned model produces a calibrated item-level diagnostic profile that reflects how strongly an item depends on different forms of modality evidence, enabling fine-grained benchmark analysis beyond aggregate accuracy. SRCoD also provides interpretable attribution of the diagnosis results. Extensive experiments under two complementary protocols, i.e., benchmark-refinement oriented validation and human-labeled item validation, demonstrate that SRCoD outperforms state-of-the-art benchmark diagnosis methods. Our code and data are available at <https://github.com/Q-17/SRCoD>.

1 Introduction

Multimodal benchmarks are central to evaluating the cross-modal reasoning ability of vision-language models (VLMs) [1–3]. These benchmarks are in essence collections of *items*, where each *item* corresponds to a complete problem instance, usually comprising multimodal inputs, a question, and a ground-truth answer [4–8]. Such benchmarks are designed to evaluate whether VLM can understand and reason across modalities [9, 10], and have inspired the development of state-of-the-art VLM models. However, recently, the quality of these benchmarks, especially whether the items are all useful for evaluating models' cross-modal inference ability, has attracted increasing attention. Early efforts have revealed that existing multimodal benchmarks often contain shortcut items that can be answered from a single modality or from answer-option regularities alone [11, 12]. Therefore, *item-level multimodal benchmark diagnosis*, which aims to identify what modality evidence each item actually requires and whether it provides a valid test of the intended cross-modal reasoning ability of downstream models, has become an important topic.

*Corresponding author.

Existing diagnostic methods can be categorized into two groups: methods based on *intrinsic item structure* and methods based on *extrinsic model responses*. Intrinsic-structure based methods [13, 14] characterize how answer-relevant evidence is distributed across modalities within the item itself, often through information-theoretic decomposition frameworks such as partial information decomposition (PID) [15]. However, item structure alone cannot reflect how downstream models actually solve the item, as an item may contain cross-modal evidence that models would not use. For example, a sidewalk-color question may contain the necessary visual evidence, yet its question and answer options can strongly favor “gray” even without the image. Extrinsic-response based methods, on the other hand, evaluate the quality of items from model behavior under different modality conditions, using item response theory (IRT) [16], multidimensional IRT [17], or multimodal variants such as M3IRT [18]. These methods first select a pool of downstream VLMs, collect their responses to each item under different modality conditions, and evaluate the multimodal benchmark from the resulting response matrix. While this grounds the diagnosis in actual model behavior, it discards the intrinsic structure of items and makes the assessment solely based on the chosen model pool. A narrow or imbalanced pool may yield response matrix that reflect model-specific strengths rather than the evidence an item actually requires. For example, identifying the resting-potential points in an action-potential graph requires both the textual question and the numbered graph. Yet failures by models with limited chart-understanding abilities may not reflect the item’s intrinsic evidence requirements.

We therefore argue that reliable multimodal benchmark diagnosis requires a *structure-response co-calibrated view*. Intrinsic item structure provides a structural signal about what evidence an item contains, while extrinsic model responses provide a behavioral signal about how the item functions in evaluation. We instantiate this view with SRCoD, a **Structure-Response Co-calibrated Diagnosis** framework for item-level multimodal benchmark diagnosis. SRCoD first derives a structural prior from sample-wise multimodal interaction signals, summarizing how answer-relevant evidence is distributed across modalities within the item. It then uses this prior to anchor a modality-conditioned IRT model fitted to extrinsic model responses under different modality settings. The resulting diagnostic profile therefore provides an attribution-style interpretation of item behavior, with each coordinate corresponding to one evidence factor. In this way, SRCoD preserves the intrinsic structure of the item while correcting it using empirical evaluation behavior.

To comprehensively assess item-level multimodal benchmark diagnosis baselines, we use two complementary validation protocols: *benchmark-refinement oriented validation* and *human-labeled item validation*. Following prior extrinsic-response work [18], we conduct benchmark-refinement oriented validation using model-ranking preservation as the validation criterion. In addition, we introduce human-labeled item validation and construct a diagnostic validation set across ScienceQA [4], MMMU [5], and SEED-Bench [6], which directly verifies whether each method recovers the dominant evidence factor underlying each item. Unlike prior intrinsic-structure work that uses human judgments mainly to assess estimator plausibility [14], our protocol collects controlled human labels indicating whether an item can be solved from the visual input alone, the textual input alone, either modality, or only their combination. Across both protocols, SRCoD outperforms intrinsic-structure and extrinsic-response baselines, with ablations confirming the benefits of structure-response co-calibration. In this paper, our contributions are threefold:

- We propose a *structure-response co-calibrated view* for item-level multimodal benchmark diagnosis, and instantiate it with SRCoD, a structure-anchored multimodal IRT framework that combines intrinsic item structure with extrinsic model responses.
- SRCoD provides interpretable attribution of the diagnosis results, indicating whether the multimodal features of an item are synergistic, redundant, visual-dominant or textual-dominant.
- We validate SRCoD under two complementary protocols, i.e., *benchmark-refinement oriented validation* and *human-labeled item validation*. Across both protocols, SRCoD outperforms state-of-the-art evaluation methods.

2 Related Work

Intrinsic-structure methods. Intrinsic-structure methods analyze how answer-relevant evidence is distributed across modalities within an item. A representative line of work uses Partial Information

Decomposition (PID) [15] to decompose multimodal information into redundant, unique, and synergistic components. Early PID-based analyses [13, 19] typically operate at the dataset-distribution level, while recent sample-wise estimators such as LSMI [14] make it possible to estimate item-level multimodal interaction structure. On real-world data, these methods often use human judgments to assess the plausibility of estimated interaction signals. However, such judgments are not controlled item-level evidence-sufficiency labels, and therefore do not directly validate whether the inferred modality-dependence profile of each benchmark item is correct.

Extrinsic-response methods. Extrinsic-response methods infer benchmark properties from model behavior. A common strategy is to evaluate models under different modality conditions and model the resulting response patterns with Item Response Theory (IRT) [16] or multidimensional IRT (MIRT) [17], treating models as subjects and benchmark examples as items. Recent multimodal extensions further decompose responses into visual, textual, and joint factors [18]. These methods are often validated through benchmark refinement protocols, such as selecting compact subsets that preserve model rankings or avoid semi-synthetic low-quality items. Such protocols demonstrate downstream validation utility, but they do not provide ground-truth item-level diagnosis labels for naturally occurring benchmark questions.

Positioning of this work. Our work addresses both the methodological and evaluation gaps above. SRCoD connects intrinsic item structure and extrinsic model responses through structure-response co-calibration, using structural interaction signals as soft anchors for modality-conditioned IRT and calibrating them with observed model responses. We further introduce a controlled human validation protocol for assessing evidence sufficiency, which provides item-level diagnostic labels for directly validating modality-dependence.

3 Method

We develop SRCoD, a structure-anchored multimodal IRT framework that instantiates structure-response co-calibration for item-level multimodal benchmark diagnosis. The intrinsic item structure summarizes how answer-relevant evidence is distributed across modalities within the item itself, while the extrinsic model responses capture how a pool of VLMs behaves under different modality conditions. The resulting modality-dependence profile summarizes four forms of item-level evidence dependence: redundant, visual-dominant, textual-dominant, and synergistic. SRCoD first derives a structural prior from sample-wise multimodal interaction signals, and then uses this prior to anchor a modality-conditioned IRT model fitted to extrinsic model responses. The fitted model produces a response-calibrated modality-dependence profile, which serves as the final diagnostic representation of the item. Since each coordinate corresponds to a distinct evidence factor, the profile also provides an interpretable attribution of item behavior.

3.1 Modality-Conditioned Response Model

Let $\mathcal{D} = \{(v_j, t_j, y_j)\}_{j=1}^J$ denote a multimodal benchmark, where v_j , t_j , and y_j are the visual input, textual input, and ground-truth answer of item j . For model i , item j , and modality condition $c = (c_v, c_t) \in \{0, 1\}^2$, we observe a binary response $r_{ijc} \in \{0, 1\}$, indicating whether the model answers correctly under condition c .

Following modality-conditioned multimodal IRT [18], we use four response coordinates corresponding to a base coordinate, a visual coordinate, a textual coordinate, and a joint visual-textual coordinate. For condition c , the active coordinates are encoded as $\phi(c) = (1, c_v, c_t, c_v c_t)^\top$. Each model i has an ability vector $\theta_i \in \mathbb{R}^4$, and each item j has discrimination and difficulty vectors $\mathbf{a}_j, \mathbf{b}_j \in \mathbb{R}^4$. The probability of a correct response is modeled as

$$P_{ijc} = P(r_{ijc} = 1) = \sigma(\phi(c)^\top (\mathbf{a}_j \odot \theta_i - \mathbf{b}_j)), \quad (1)$$

where $\odot$ denotes element-wise multiplication. This likelihood captures how model correctness changes when different modality coordinates are activated.

If $\mathbf{a}_j$ and $\mathbf{b}_j$ were learned solely from responses, the resulting diagnosis would be response-only and could be sensitive to the diagnostic model pool. The key distinction of SRCoD is that these item parameters are anchored by intrinsic item structure before being calibrated by extrinsic model responses.

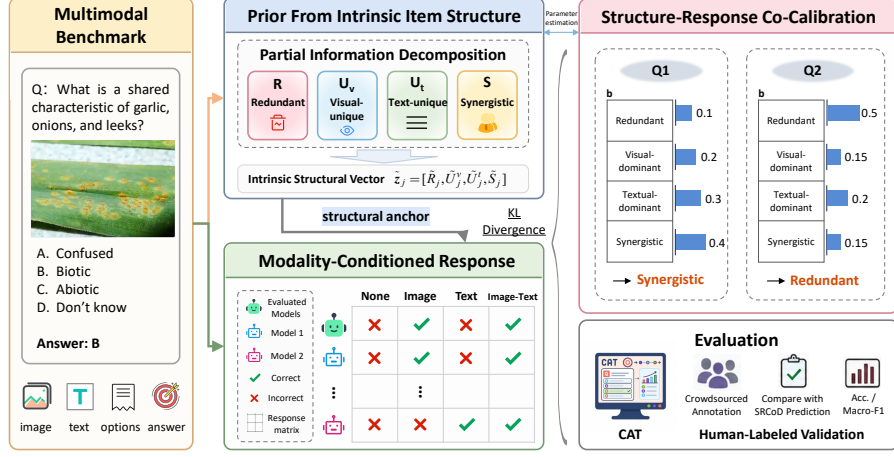


Figure 1: Overview of SRCoD. SRCoD diagnoses item-level modality dependence by co-calibrating intrinsic item structure and extrinsic model responses. It first derives a structural prior from partial information decomposition, then anchors a modality-conditioned response model fitted to VLM responses under controlled modality conditions.

3.2 Prior From Intrinsic Item Structure

Response-only diagnosis can conflate modality dependence with empirical difficulty. To reduce this confounding, SRCoD introduces a structural prior derived from intrinsic item structure.

For each item $j = (v_j, t_j, y_j)$, we estimate a sample-wise multimodal interaction vector using the Lightweight Sample-wise Multimodal Interaction (LSMI) estimator [14]. After robust normalization, the interaction vector is written as

$$\tilde{\mathbf{z}}_j = (\tilde{R}_j, \tilde{U}_j^v, \tilde{U}_j^t, \tilde{S}_j)^\top, \quad (2)$$

where $\tilde{R}_j$ denotes redundant evidence, $\tilde{U}_j^v$ and $\tilde{U}_j^t$ denote visual-unique and textual-unique evidence, and $\tilde{S}_j$ denotes synergistic evidence.

We convert this interaction vector into structural dependence scores. First, we define modality answerability scores as $A_{j,v} = \tilde{R}_j + \tilde{U}_j^v$, $A_{j,t} = \tilde{R}_j + \tilde{U}_j^t$, and $A_{j,vt} = \tilde{R}_j + \tilde{U}_j^v + \tilde{U}_j^t + \tilde{S}_j$. The structural dependence scores are then given by

$$\boldsymbol{\psi}_j = (\min(A_{j,v}, A_{j,t}), [A_{j,v} - A_{j,t}]_+, [A_{j,t} - A_{j,v}]_+, [A_{j,vt} - \max(A_{j,v}, A_{j,t})]_+), \quad (3)$$

where $[x]_+ = \max(x, 0)$. The four coordinates correspond to redundant, visual-dominant, textual-dominant, and synergistic dependence, respectively.

The structural prior is defined as

$$\mathbf{q}_j^{\text{str}} = \text{softmax}(\boldsymbol{\psi}_j). \quad (4)$$

Because intrinsic structural estimates can be ambiguous, SRCoD does not treat this prior as a hard label. Instead, it assigns each item an entropy-based confidence weight,

$$w_j = \left[1 - \frac{H(\mathbf{q}_j^{\text{str}})}{\log 4} - \gamma \right]_+. \quad (5)$$

A near-uniform structural prior therefore has little influence, whereas a low-entropy prior contributes more strongly during structure-response co-calibration.

3.3 Structure-Anchored Item Parameters

The structural interaction coordinates and the response-side IRT coordinates are related but not identical. The structural coordinates describe how answer-relevant evidence is distributed within the item, whereas the IRT coordinates describe how model correctness changes under modality

interventions. To connect these two views, SRCoD learns a bridge from intrinsic item structure to modality-conditioned item parameters.

Let

$$\mathbf{z}_j^* = (\tilde{\mathbf{z}}_j - \boldsymbol{\mu}_z) \oslash \boldsymbol{\sigma}_z \quad (6)$$

be the standardized structural interaction vector. We project it into the response-coordinate space by

$$\mathbf{h}_j = W\mathbf{z}_j^* + \boldsymbol{\xi}, \quad \mathbf{h}_j = (h_{j,0}, h_{j,v}, h_{j,t}, h_{j,vt})^\top. \quad (7)$$

The projected vector $\mathbf{h}_j$ serves as a structural anchor for the item parameters of the modality-conditioned IRT model.

Specifically, the discrimination and difficulty parameters are parameterized as

$$a_{j,k} = \text{softplus}(\alpha_k h_{j,k} + \beta_k + \Delta_{j,k}^a) + \epsilon_0, \quad b_{j,k} = \text{softplus}(\gamma_k h_{j,k} + \eta_k + \Delta_{j,k}^b) + \epsilon_0, \quad (8)$$

for $k \in \{0, v, t, vt\}$. Here $\alpha_k, \beta_k, \gamma_k, \eta_k$ are global calibration parameters, while $\Delta_{j,k}^a$ and $\Delta_{j,k}^b$ are item-specific residuals learned from extrinsic model responses.

This parameterization implements the anchoring step in structure-response co-calibration. The anchor term injects intrinsic structural evidence into the response model, while the residual terms allow response-side calibration to revise the structural prior when observed model behavior disagrees with the structural estimate.

3.4 Response-Calibrated Modality-Dependence Profile

After fitting the structure-anchored response model, SRCoD derives a calibrated modality-dependence profile from the learned response surface. Let $\boldsymbol{\theta}$ denote the average ability vector of the diagnostic model pool. For each item j , we compute calibrated correctness probabilities

$$p_{j,s} = P(r = 1 \mid j, c_s, \boldsymbol{\theta}), \quad s \in \{0, v, t, vt\}. \quad (9)$$

For notational compactness, we write $p_s = p_{j,s}$ for a fixed item j . We define the response-side dependence scores as

$$\boldsymbol{\chi}_j = (\min(p_v, p_t) - p_0, p_v - \max(p_t, p_0), p_t - \max(p_v, p_0), p_{vt} - \max(p_v, p_t)). \quad (10)$$

The four coordinates correspond to redundant, visual-dominant, textual-dominant, and synergistic dependence, respectively. They measure shared unimodal gain over the no-modality condition, asymmetric unimodal advantages, and the extra gain of the full multimodal condition over the best unimodal condition.

The final response-calibrated modality-dependence profile is

$$\mathbf{q}_j^{\text{cal}} = \text{softmax}(\boldsymbol{\chi}_j). \quad (11)$$

Compared with raw accuracy differences across modality conditions, $\mathbf{q}_j^{\text{cal}}$ is calibrated through latent model ability and structure-anchored item parameters, reducing sensitivity to the particular composition of the diagnostic model pool. We therefore interpret $\mathbf{q}_j^{\text{cal}}$ not only as a modality-dependence profile, but also as an attribution over evidence factors that explains which source of evidence accounts for the item's observed evaluation behavior.

3.5 Structure-Response Co-Calibration Objective

SRCoD is trained by combining response-side likelihood fitting with structural-prior alignment. The final objective is

$$\mathcal{L} = - \sum_{(i,j,c) \in \Omega} [r_{ijc} \log P_{ijc} + (1 - r_{ijc}) \log(1 - P_{ijc})] + \frac{\lambda_{\text{str}}}{J} \sum_{j=1}^J w_j \text{KL}(\mathbf{q}_j^{\text{str}} \parallel \mathbf{q}_j^{\text{cal}}). \quad (12)$$

The first term calibrates the response model using observed model behavior, while the second term preserves intrinsic item structure according to the confidence weight w_j . Thus, confident structural estimates act as stronger anchors, whereas ambiguous estimates are mainly resolved by extrinsic model responses.

After training, SRCoD uses $\mathbf{q}_j^{\text{cal}}$ as the final item-level diagnostic representation. Its coordinates quantify whether the calibrated evidence for item j is primarily redundant, visual-dominant, textual-dominant, or synergistic. This yields a modality-dependence profile that preserves intrinsic item structure while being corrected by empirical response behavior.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on three widely used vision-language benchmarks: ScienceQA [4], MMMU [5], and SEED-Bench [6]. They cover science-oriented question answering, college-level multimodal reasoning, and general multimodal understanding, respectively.

Evaluation settings. We validate benchmark diagnosis methods under two complementary protocols. First, *benchmark-refinement oriented validation* follows the contaminated candidate-pool setting of prior work [18]: each pool contains both clean benchmark items and synthetic low-quality items generated by swapping image or text components across questions. A method selects a compact subset from this pool, and the subset is evaluated by how well it preserves the model ranking induced by the full clean benchmark.

Second, *human-labeled item validation* evaluates whether a method recovers the dominant evidence factor of each item. We construct a human-labeled diagnostic validation set from ScienceQA, MMMU, and SEED-Bench. Candidate items are first filtered under the original full multimodal condition to ensure human solvability, and each retained item is then annotated under controlled evidence conditions. This yields a human evidence-sufficiency reference in a four-factor diagnostic space: Redundant, Visual-dominant, Textual-dominant, or Synergistic. The validation set contains 1,000 items per benchmark and is used only for evaluation. For quantitative evaluation, we compare each method’s dominant diagnostic factor with the human reference and report accuracy and macro-F1. Full annotation details are provided in Appendix C.

Model pool and response collection. For methods using extrinsic responses, we collect modality-conditioned responses from a heterogeneous pool of VLMs, including proprietary and open-weight models from GPT-series [20], Qwen-VL [21], Gemma [22], Ministral [23], and other model families. The full model list, prompts, and implementation details are provided in Appendix D.

Compared methods. We compare SRCoD with six baselines in total. Random selects a diagnostic factor uniformly at random. Raw directly maps modality-conditioned correctness gaps to the four diagnostic factors. LSMI [14] is a structure-only baseline that infers the dominant evidence factor from sample-wise multimodal interaction signals. IRT [16] is a standard psychometric baseline based on item difficulty, while MIRT [17] extends IRT with modality-conditioned response factors. M3IRT [18] is a stronger multimodal IRT baseline that models modality-specific item behavior from extrinsic responses. In contrast, SRCoD incorporates intrinsic structural signals as soft priors and infers the dominant evidence factor from its learned response-calibrated profile.

4.2 Main Results

Benchmark-refinement oriented validation. We first conduct benchmark-refinement oriented validation to assess whether diagnostic scores can support reliable benchmark-based model comparison. In this setting, no item-level diagnostic labels are used. Instead, following the contaminated candidate-pool protocol of prior work [18], we evaluate whether a diagnosis method can select compact subsets that preserve the model-ranking conclusions of the full clean benchmark.

For each benchmark, we first compute a reference model ranking using the full clean set. We then construct a contaminated candidate pool by combining the clean items with an equal number of synthetic low-quality items. Following prior work [18], the low-quality items are generated by swapping image or text components across different questions, producing shortcut-prone or unsolvable examples while preserving the benchmark format. The resulting candidate pool contains 50% clean items and 50% synthetic low-quality items.

We compare Random selection, IRT, M3IRT, and SRCoD. For each method, we select subsets with sizes corresponding to 5%, 10%, 15%, 20%, and 25% of the clean set size. The selected subset is then used to compute model accuracies and induce a subset-based model ranking. We measure ranking preservation by the Spearman rank correlation between this subset-induced ranking and the reference ranking obtained from the full clean benchmark:

$$\rho = \text{Spearman}(\text{rank}_{\text{clean}}, \text{rank}_{\text{subset}}). \quad (13)$$

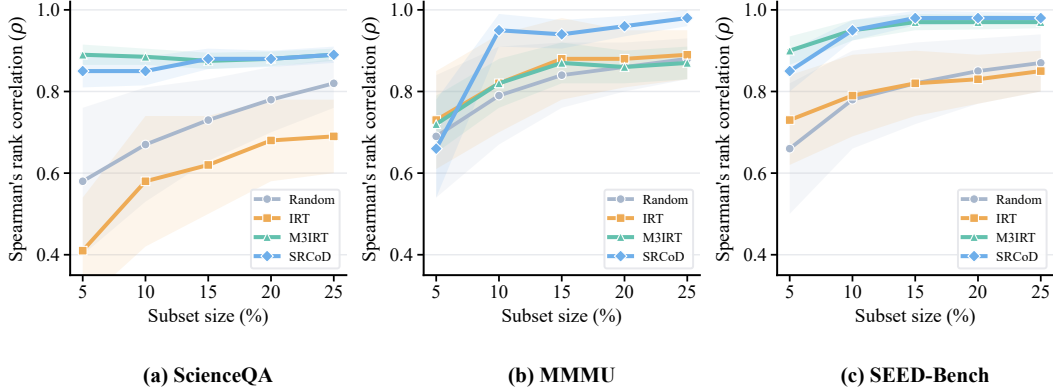


Figure 2: Benchmark-refinement oriented validation under contaminated candidate pools. Each method selects compact subsets from candidate pools containing equal numbers of clean benchmark items and synthetic low-quality items. We report the Spearman rank correlation between the model ranking induced by the selected subset and the ranking induced by the full clean benchmark.

A higher ρ indicates that the selected subset better preserves the evaluation conclusion of the full clean benchmark. For each subset size, we repeat item-pool resampling 20 times and report the mean and standard deviation of Spearman rank correlation.

Figure 2 shows that SRCoD achieves strong benchmark-refinement oriented validation performance under contaminated candidate pools. On ScienceQA, SRCoD maintains high rank correlation across all subset sizes and is consistently stronger than IRT and Random, while remaining competitive with M3IRT. On MMMU, SRCoD achieves the best performance for subset sizes of 10% and above, with rank correlation approaching the full clean benchmark ranking even with relatively small subsets. On SEED-Bench, SRCoD and M3IRT both perform strongly, but SRCoD reaches near-ceiling rank correlation from 10% onward and remains substantially better than IRT and Random. Overall, these results indicate that the diagnostic profiles produced by SRCoD can identify items that preserve reliable model-comparison conclusions even when the candidate pool is contaminated.

Human-labeled item validation. We further evaluate whether the diagnostic profile identifies the correct dominant evidence factor using the human-labeled validation set. The comparison among baselines highlights the complementary roles of intrinsic item structure and extrinsic model responses. The structure-only LSMI baseline performs above random assignment in accuracy but obtains relatively low macro-F1, indicating that intrinsic structural signals capture some dominant dependence patterns but are insufficient for balanced diagnosis across all categories. Extrinsic-response baselines generally perform better, showing that modality-conditioned responses provide useful empirical evidence for inferring the dominant evidence factor. However, their performance still lags behind SRCoD, suggesting that response evidence alone does not fully resolve fine-grained modality-dependence diagnosis.

SRCoD consistently outperforms the baselines by combining the two sources of evidence. The structural prior preserves intrinsic modality-dependence information, while response-side calibration adjusts this prior using observed modality-conditioned responses. These results support the central motivation of SRCoD: reliable multimodal benchmark diagnosis benefits from structure-response co-calibration rather than relying solely on intrinsic item structure or extrinsic model responses.

4.3 Ablation Study

We conduct an ablation study to examine two key components of SRCoD: entropy-based confidence weighting and profile-level structural alignment. The *w/o entropy* variant removes the entropy-based weight w_j , treating all structural priors as equally reliable. The *w/o struct.* variant removes the structural alignment loss $\mathcal{L}_{\text{str}}$, while keeping the response-side model and structure-anchored item parameterization unchanged.

Table 1: Human-labeled item validation. Accuracy and macro-F1 measure agreement with human evidence-sufficiency references in the four-factor diagnostic space. Bold and underline denote the best and second-best results, respectively.

Method	ScienceQA		MMMU		SEED-Bench		Average	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
Random	0.251	0.233	0.249	0.241	0.250	0.242	0.250	0.239
Raw	0.353	0.334	<u>0.503</u>	0.435	0.388	0.355	0.415	0.375
LSMI	0.464	0.255	0.265	0.171	0.309	0.189	0.346	0.205
IRT	0.507	0.476	0.384	0.382	0.494	0.474	0.462	0.444
MIRT	0.506	0.476	0.371	0.370	0.494	0.473	0.457	0.440
M3IRT	0.570	0.538	0.486	0.484	0.618	0.603	0.558	0.542
SRCoD	0.607	0.588	0.548	0.538	0.685	0.673	0.613	0.600

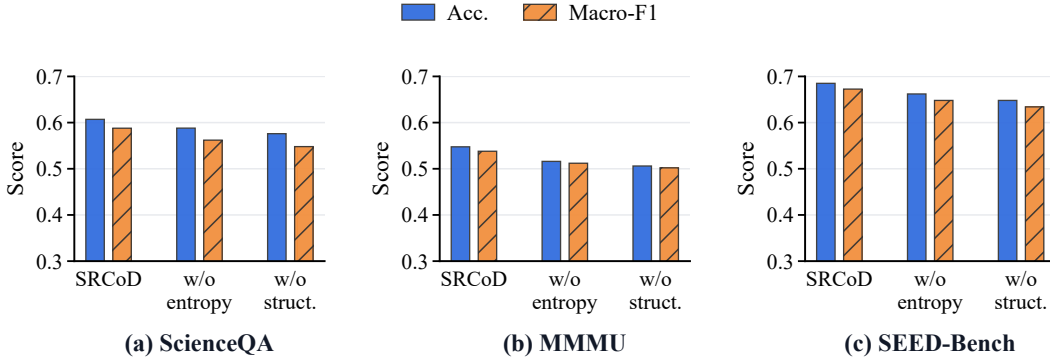


Figure 3: Ablation study of SRCoD. We compare the full model with variants that remove entropy-based confidence weighting or the profile-level structural alignment loss.

Figure 3 shows that the full SRCoD model consistently achieves the best accuracy and macro-F1 across the three benchmarks. Removing entropy weighting degrades performance, indicating that uncertain structural priors should not be enforced as strongly as confident ones. Entropy weighting therefore improves robustness by down-weighting ambiguous intrinsic structural signals. Removing profile-level structural alignment also reduces performance. This suggests that intrinsic structural signals are most effective when they both anchor item parameters and regularize the final response-calibrated diagnostic profile.

Overall, the ablation results show that the two components play complementary roles: entropy weighting controls the reliability of structural priors, while profile-level alignment transfers intrinsic structural information to the final diagnostic profile.

4.4 Robustness to Model-Pool Subsampling

A reliable benchmark diagnosis method should produce stable item-level conclusions under different diagnostic model pools. This is important for multimodal benchmark auditing, because collecting modality-conditioned responses from a large pool of VLMs can be costly. We therefore validate robustness when only a subset of diagnostic models is available.

For each benchmark, we randomly sample $n \in \{4, 6, 8, 10, 12\}$ models from the full diagnostic model pool. For each sampled pool, we fit the diagnosis method using only the selected models and validate item-level predictions on the same human-labeled validation set. We repeat this procedure 10 times for each pool size and report performance retention relative to the full-pool setting:

$$\text{Retention}^{(s)}(n) = \frac{\text{Score}^{(s)}(n)}{\text{Score}(\text{all})}, \quad (14)$$

where s indexes the subsampling trial, and Score denotes either accuracy or macro-F1. Figure 4 reports the mean retention over 10 trials, with shaded regions denoting one standard deviation.

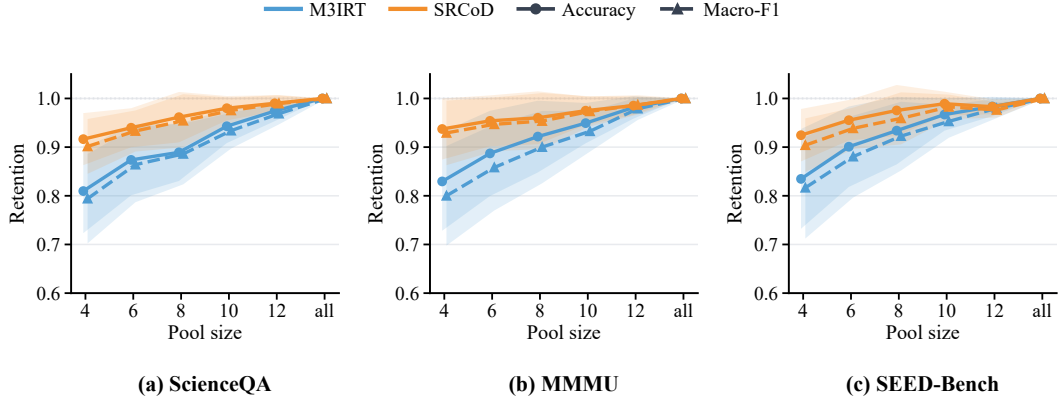


Figure 4: Robustness to model-pool subsampling. For each pool size, we randomly sample diagnostic models from the full model pool and repeat the experiment 10 times. We report performance retention relative to the full-pool setting. Shaded regions denote one standard deviation.

Figure 4 shows that SRCoD is more robust than M3IRT under small diagnostic model pools. When only four or six models are used, M3IRT shows a clear drop in accuracy and macro-F1 retention, whereas SRCoD stays closer to its full-pool performance across the three benchmarks. SRCoD also exhibits smaller variation across subsampling trials, indicating that intrinsic structural anchoring reduces sensitivity to the particular choice of diagnostic models.

Overall, these results provide additional evidence for structure-anchored diagnosis. By combining extrinsic model responses with intrinsic item structure, SRCoD produces more stable modality-dependence estimates when the diagnostic model pool is limited.

We further validate the behavioral generalization of SRCoD on held-out model families in Appendix F.4, where the predicted diagnostic categories consistently exhibit their expected modality-dependent response patterns.

4.5 Qualitative Diagnostic Example

Figure 5 presents a representative example of the diagnostic output produced by SRCoD. The question asks for the capital of Wyoming, while the accompanying map highlights Wyoming in dark green. The correct answer is *Cheyenne*. From the textual input alone, a model can identify the state directly from the question and retrieve its capital using geographic knowledge, without consulting the image. Meanwhile, the visual input provides an alternative way to identify the state through the highlighted region on the map. Consequently, both modalities convey sufficient information to establish the geographic entity needed to answer the question. Based on this overlapping evidence, SRCoD classifies the item as *Redundant*. This example demonstrates that SRCoD provides an interpretable diagnostic result by identifying how the available evidence contributes to answerability.

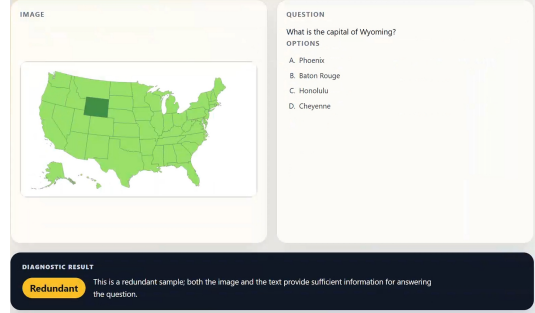


Figure 5: Qualitative example of the diagnostic output produced by SRCoD. The item is diagnosed as *Redundant*, indicating that both the image and the text provide sufficient evidence for answering the question.

5 Conclusion

This paper studied item-level multimodal benchmark diagnosis. We proposed SRCoD, a structure-anchored multimodal IRT framework that co-calibrates intrinsic item structure with extrinsic model responses. By using sample-wise multimodal interaction signals as structural priors and calibrating them with modality-conditioned response behavior, SRCoD produces fine-grained diagnostic profiles

for redundant, visual-dominant, textual-dominant, and synergistic items. We also introduced a controlled human validation protocol and constructed a diagnostic validation set across ScienceQA, MMMU, and SEED-Bench. Experiments show that SRCoD outperforms intrinsic-structure and extrinsic-response baselines in item-level diagnosis, remains more stable under limited model pools, and preserves model rankings when selecting compact benchmark subsets. Future work may extend this framework to open-ended tasks and broader multimodal settings beyond vision and language.

Acknowledgments and Disclosure of Funding

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. U23A20296, 62272469 and 62302513), and the Science and Technology Innovation Program of Hunan Province (2026RC3144).

References

- [1] Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5625–5644, 2024.
- [2] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling, 2024.
- [3] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 87874–87907, 2024.
- [4] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.
- [5] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL <https://doi.org/10.1109/CVPR52733.2024.00913>.
- [6] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13299–13308. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01263. URL <https://doi.org/10.1109/CVPR52733.2024.01263>.
- [7] 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. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, Lecture Notes in Computer Science, pages 216–233. Springer, 2024. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

- [8] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? EMMA: an enhanced multimodal reasoning benchmark. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL <https://proceedings.mlr.press/v267/hao25a.html>.
- [9] Shiqi Zhang, Weixin Zeng, Zhen Tan, Weidong Xiao, and Xiang Zhao. HIM2A: hierarchical interactive multi-modal entity alignment with semantic augmentation. *Adv. Eng. Informatics*, 70:104136, 2026. doi: 10.1016/J.AEI.2025.104136. URL <https://doi.org/10.1016/j.aei.2025.104136>.
- [10] Shiqi Zhang, Weixin Zeng, Ziheng Zhang, Wenzhe Hou, Weidong Xiao, and Xiang Zhao. Comco: Consistency-aware multi-agent coordination for zero-shot cross-modal entity matching. In Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang, editors, *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, Melbourne VIC Australia, July 20-24, 2026*, pages 2398–2409. ACM, 2026. doi: 10.1145/3805712.3809543. URL <https://doi.org/10.1145/3805712.3809543>.
- [11] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414, 2019. doi: 10.1007/S11263-018-1116-0. URL <https://doi.org/10.1007/s11263-018-1116-0>.
- [12] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.
- [13] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas B. Allen, Randy Auerbach, Faisal Mahmood, Russ Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/575286a73f238b6516ce0467d67eadb2-Abstract-Conference.html.
- [14] Zequn Yang, Hongfa Wang, and Di Hu. Efficient quantification of multimodal interaction at sample level. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL <https://proceedings.mlr.press/v267/yang25aj.html>.
- [15] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010. URL <https://arxiv.org/abs/1004.2515>.
- [16] Xitao Fan. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, 58(3):357–381, 1998.
- [17] Mark D Reckase. 18 multidimensional item response theory. *Handbook of statistics*, 26: 607–642, 2006.

- [18] Shunki Uebayashi, Kento Masui, Kyohei Atarashi, Han Bao, Hisashi Kashima, Naoto Inoue, Mayu Otani, and Koh Takeuchi. Evaluating cross-modal reasoning ability and problem characteristics with multimodal item response theory, 2026. URL <https://arxiv.org/abs/2603.02663>.
- [19] Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina K. Ramirez, Gregory Heller, Severine Durand, Shawn R. Olsen, and Stefan Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ec0bff8bf4b11e36f874790046dfdb65-Abstract-Conference.html.
- [20] OpenAI. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- [21] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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. Qwen2.5-vl technical report, 2025. URL <https://arxiv.org/abs/2502.13923>.
- [22] Gemma Team. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.
- [23] Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyt, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Praveesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL <https://arxiv.org/abs/2601.08584>.
- [24] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.
- [25] Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL <https://arxiv.org/abs/2312.11805>.
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

- [27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

A Limitation

This work has three main limitations. First, our experiments focus on multiple-choice vision-language benchmarks. Although ScienceQA, MMMU, and SEED-Bench cover diverse domains and reasoning types, SRCoD has not yet been validated on open-ended generation tasks, interactive multimodal evaluation, or modalities beyond vision and language.

Second, SRCoD relies on a diagnostic pool of VLMs to collect modality-conditioned responses. While structural anchoring improves robustness, the resulting profiles remain sensitive to the composition and capabilities of the selected models. For example, limited understanding of molecular diagrams may lead to overestimated cross-modal synergy, while difficulties in interpreting contingency tables may cause visual contributions to be underestimated.

Third, both the intrinsic structural prior and the human validation labels involve uncertainty. Sample-wise multimodal interaction estimates are approximations of item-level evidence structure and may be affected by representation quality. Meanwhile, controlled single-modality human labels can depend on annotator background knowledge and item ambiguity.

These limitations suggest that SRCoD provides estimates of modality dependence rather than definitive judgments of human evidence sufficiency. Future work may explore broader evaluation settings, more representative diagnostic model pools, and graded evidence-sufficiency annotations.

B Preliminaries

B.1 Partial Information Decomposition

Partial Information Decomposition (PID) provides a framework for analyzing how multiple information sources contribute to a target variable. Given two sources X^v and X^t , and a target variable Y , PID decomposes the target-relevant information into four components:

$$(R, U_v, U_t, S), \quad (15)$$

where R denotes redundant information shared by both sources, U_v and U_t denote source-specific information uniquely provided by X^v and X^t , respectively, and S denotes synergistic information that is available only from the joint use of both sources.

In the two-source case, these components satisfy

$$I(X^v; Y) = R + U_v, \quad I(X^t; Y) = R + U_t, \quad I(X^v, X^t; Y) = R + U_v + U_t + S. \quad (16)$$

In vision-language evaluation, the two sources naturally correspond to the visual input and the textual input, while the target corresponds to the answer. This makes PID a useful conceptual tool for distinguishing whether answer-relevant information is available redundantly from either modality, uniquely from one modality, or only through their interaction.

B.2 Item Response Theory and Multimodal Extensions

Item Response Theory (IRT) models binary responses as a function of subject ability and item properties. In benchmark evaluation, models can be treated as subjects and benchmark examples as items. The standard two-parameter logistic model is

$$P(r_{ij} = 1) = \sigma(a_j(\theta_i - b_j)), \quad (17)$$

where θ_i denotes the ability of subject i , and a_j and b_j denote the discrimination and difficulty of item j .

Multidimensional IRT (MIRT) generalizes IRT by representing ability and discrimination in a latent vector space:

$$P(r_{ij} = 1) = \sigma(\mathbf{a}_j^\top \boldsymbol{\theta}_i - b_j), \quad (18)$$

where $\boldsymbol{\theta}_i$ is a multidimensional ability vector and $\mathbf{a}_j$ is a multidimensional discrimination vector.

For vision-language evaluation, multimodal IRT further conditions the response process on modality availability. It decomposes model ability and item properties into modality-related coordinates, such as base, visual, textual, and joint visual-textual components. SRCoD adopts this modality-conditioned response modeling principle, but differs from response-only approaches by anchoring item parameters with intrinsic item structure.

C Human Annotation Protocol

This section describes the construction of the human-labeled diagnostic validation set used in Section 4. The validation set is designed to provide an independent evaluation signal for item-level multimodal benchmark diagnosis. It is used only for evaluation and is not used to train SRCoD, tune its hyperparameters, or fit any baseline method.

C.1 Candidate Items

We construct the validation set from three benchmarks: ScienceQA [4], MMMU [5], and SEED-Bench [6]. Candidate items are sampled from their multiple-choice vision-language question answering subsets. Each item consists of a visual input, a textual question, a set of answer options, and the ground-truth answer.

To ensure that diagnostic labels reflect modality dependence rather than item invalidity, we first filter candidate items under the original full multimodal condition. In this condition, annotators are given the original image, question, and answer options. Each candidate item is answered by $M = 3$ annotators. Let $r_{mj}^f \in \{0, 1\}$ denote whether annotator m answers item j correctly under the full multimodal condition. We retain item j only if a majority of annotators answer it correctly:

$$s_f(j) = \mathbb{I} \left[\sum_{m=1}^M r_{mj}^f \geq \left\lceil \frac{M}{2} \right\rceil \right]. \quad (19)$$

Items with $s_f(j) = 1$ form the retained pool. From this pool, we sample 1,000 items per benchmark to form the diagnostic validation set.

C.2 Controlled Evidence Conditions

For each retained item, we collect human answers under two controlled evidence conditions. These conditions are designed to test whether a single modality provides sufficient evidence for solving the item. To reduce carryover effects, an annotator is assigned at most one controlled evidence condition for a given item, and item order is randomized within each annotation batch. Annotators are instructed to select exactly one option for each item and are not allowed to access external resources.

Visual-conditioned setting. Annotators are provided with the visual input and answer options. When necessary, we provide a minimal task instruction that preserves the task format while removing answer-bearing textual evidence. For example, the instruction may specify that the annotator should select the option best supported by the image, but it does not include object names, relations, quantities, or other cues that would reveal the answer without visual evidence. This condition is used to test whether the item can be solved primarily from visual evidence.

Textual-conditioned setting. Annotators are provided with the textual question and answer options, but not the visual input. This condition is used to test whether the item can be solved primarily from textual cues or language priors.

Each retained item is answered by $M = 5$ annotators under each controlled evidence condition. The annotator index is condition-specific and does not imply that the same annotator sees the same item under both controlled conditions.

C.3 Label Assignment

Let $r_{mj}^v \in \{0, 1\}$ denote whether annotator m answers item j correctly under the visual-conditioned setting, and let $r_{mj}^t \in \{0, 1\}$ denote whether annotator m answers item j correctly under the textual-conditioned setting. We define two binary sufficiency indicators:

$$s_v(j) = \mathbb{I} \left[\sum_{m=1}^M r_{mj}^v \geq \left\lceil \frac{M}{2} \right\rceil \right], \quad s_t(j) = \mathbb{I} \left[\sum_{m=1}^M r_{mj}^t \geq \left\lceil \frac{M}{2} \right\rceil \right]. \quad (20)$$

A controlled evidence condition is therefore considered sufficient if a majority of annotators answer the item correctly under that condition.

Each item is assigned one of four modality-dependence labels:

$$y_j = \begin{cases} \text{Redundant,} & s_v(j) = 1, s_t(j) = 1, \\ \text{Visual-dominant,} & s_v(j) = 1, s_t(j) = 0, \\ \text{Textual-dominant,} & s_v(j) = 0, s_t(j) = 1, \\ \text{Synergistic,} & s_v(j) = 0, s_t(j) = 0. \end{cases} \quad (21)$$

The Redundant label indicates that either controlled evidence condition is sufficient for human annotators to answer the item. The Visual-dominant and Textual-dominant labels indicate that the item is primarily supported by visual or textual evidence, respectively. The Synergistic label indicates that neither controlled evidence condition is sufficient, while the item remains human-solvable under the original full multimodal condition.

C.4 Annotation Reliability

We report inter-annotator reliability using Fleiss’ Kappa [24] to quantify the consistency of human judgments. For the full multimodal filtering stage, Fleiss’ Kappa is 0.76, indicating substantial agreement among annotators. For the two controlled evidence conditions, Fleiss’ Kappa is 0.56 for the visual-conditioned setting and 0.63 for the textual-conditioned setting. These values indicate moderate agreement, reflecting the greater ambiguity of judging item answerability from a single modality. Overall, the annotation results suggest that the retained full-condition items are consistently human-solvable, while the controlled single-modality labels provide reliable but non-trivial diagnostic signals.

C.5 Label Distribution of the Diagnostic Validation Set

Figure 6 summarizes the modality-dependence label distribution of the human-labeled diagnostic validation set. The proportions of Redundant, Visual-dominant, Textual-dominant, and Synergistic items vary across ScienceQA, MMMU, and SEED-Bench, indicating that different benchmarks contain different mixtures of shortcut-prone and cross-modally dependent items. This variation further motivates item-level benchmark diagnosis rather than assuming a fixed dataset-level modality structure.

D Details of Experimental Settings

D.1 Evaluation Metric

We evaluate all methods against the human-labeled diagnostic validation set. Each item j has a human annotation label $y_j \in \mathcal{Y}$, where

$$\mathcal{Y} = \{\text{Redundant, Visual-dominant, Textual-dominant, Synergistic}\}. \quad (22)$$

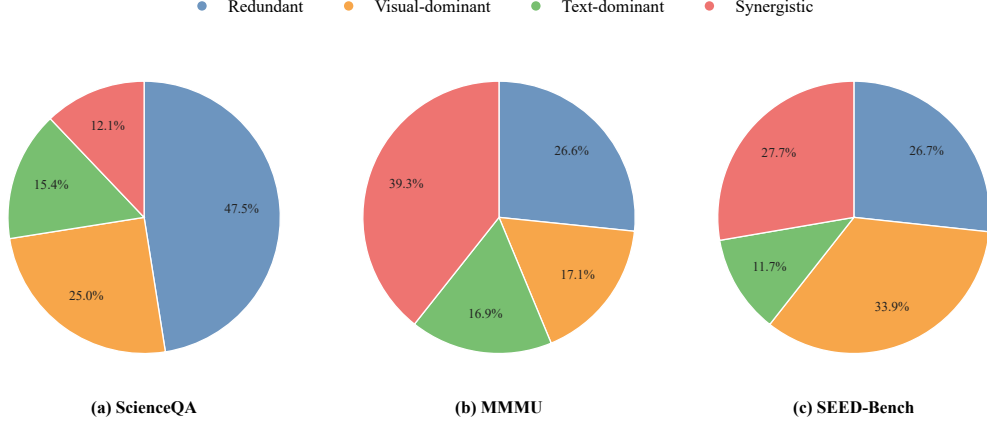


Figure 6: Distribution of human-annotated modality-dependence labels across ScienceQA, MMMU, and SEED-Bench. Each pie chart shows the proportion of Redundant, Visual-dominant, Textual-dominant, and Synergistic items in the corresponding benchmark.

For methods that output a four-dimensional modality-dependence profile, we predict the label by taking the coordinate with the largest score:

$$\hat{y}_j = \arg \max_{c \in \mathcal{Y}} q_{j,c}. \quad (23)$$

For methods that output modality-conditioned scores or probabilities, we first map their outputs to the same four-factor diagnostic space and then apply the same argmax rule.

We report accuracy and macro-F1 as the main evaluation metrics. Accuracy measures the overall fraction of correctly predicted labels:

$$\text{Acc} = \frac{1}{N} \sum_{j=1}^N \mathbb{I}[\hat{y}_j = y_j], \quad (24)$$

where N is the number of items in the diagnostic validation set.

To compute macro-F1, we first compute precision, recall, and F1 for each class $c \in \mathcal{Y}$. Let TP_c , FP_c , and FN_c denote the number of true positives, false positives, and false negatives for class c . The class-wise precision and recall are

$$\text{Prec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \quad \text{Rec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}. \quad (25)$$

The F1 score for class c is then

$$\text{F1}_c = \frac{2 \text{Prec}_c \text{Rec}_c}{\text{Prec}_c + \text{Rec}_c}. \quad (26)$$

Finally, macro-F1 is computed by averaging the class-wise F1 scores over the four diagnostic categories:

$$\text{Macro-F1} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \text{F1}_c. \quad (27)$$

We use macro-F1 in addition to accuracy because the human-labeled diagnostic set is not perfectly class-balanced. Macro-F1 gives equal weight to Redundant, Visual-dominant, Textual-dominant, and Synergistic items, making it more informative for evaluating whether a method performs consistently across all modality-dependence categories.

D.2 Computational Resources

The computational resources used in this study are summarized in Table 2. All modality-conditioned evaluations estimation are conducted on the same computing environment. For each benchmark, we collect model responses under the required modality conditions and then fit the corresponding diagnostic models.

Table 2: Computational resources used for experiments.

Component	Specification
CPU	12th Gen Intel(R) Core(TM) i7-12700KF @ 3.60 GHz
RAM	32 GB
GPU	NVIDIA GeForce RTX 4090, 24 GB
Python version	3.10.1
PyTorch version	2.7.0
CUDA version	12.6

D.3 Datasets

We use three vision-language benchmarks in our experiments.

ScienceQA [4] is a science-oriented multimodal question answering benchmark. It contains multiple-choice questions with textual contexts and, when available, visual inputs. The license for the ScienceQA dataset is Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International.

MMMU [5] is a massive multi-discipline multimodal understanding and reasoning benchmark designed to evaluate college-level reasoning ability across diverse domains, including science, medicine, engineering, business, humanities, and art. The license for the MMMU dataset is Apache License 2.0.

SEED-Bench [6] is a large-scale benchmark for evaluating general multimodal understanding abilities using multiple-choice questions. We use its vision-language question answering items in our experiments. The official release materials indicate that SEED-Bench is released under Apache License 2.0, while the Hugging Face dataset card lists Attribution-NonCommercial 4.0 International and notes that some underlying visual assets are subject to their original dataset licenses.

D.4 Models

We use a diverse pool of multimodal large language models as diagnostic respondents. As summarized in Table 3, the pool includes both closed-source API models and open-weight models, covering different model families, parameter scales, and deployment settings. Closed models provide strong contemporary VLM baselines, while open-weight models increase transparency and make it possible to examine performance variation across model sizes. This heterogeneous model pool is used to collect modality-conditioned responses for fitting the response-side components of SRCoD and the compared response-based baselines. No model in the pool is fine-tuned during our experiments; all models are evaluated under the same benchmark prompts and modality conditions.

Table 3: Overview of AI Models Used

Model Name	Type	License or Terms
gpt-4.1 [20]	Closed	OpenAI Terms of Use
gpt-4.1-nano [20]	Closed	OpenAI Terms of Use
gpt-4o-mini [20]	Closed	OpenAI Terms of Use
gemini-3.1-pro-preview [25]	Closed	Gemini API Additional Terms of Service
gemini-3.1-flash-lite [25]	Closed	Gemini API Additional Terms of Service
gemini-2.5-flash-lite [25]	Closed	Gemini API Additional Terms of Service
llava:13b [26]	Open-weight	Llama 2 Community License Agreement
llava:7b [26]	Open-weight	Apache 2.0
gemma3:12b [22]	Open-weight	Gemma Terms of Use
gemma3:4b [22]	Open-weight	Gemma Terms of Use
ministral-3:14b [23]	Open-weight	Apache 2.0
ministral-3:8b [23]	Open-weight	Apache 2.0
qwen2.5vl:7b [21]	Open-weight	Apache 2.0
qwen2.5vl:3b [21]	Open-weight	Apache 2.0

Table 4: Input information provided under each modality condition.

Condition	Image	Question text	Answer options
Full multimodal	Yes	Yes	Yes
Visual-conditioned	Yes	No	Yes
Textual-conditioned	No	Yes	Yes
No-modality	No	No	Yes

D.5 Prompts

We use standardized multiple-choice prompts to collect model responses under different modality conditions. For all models and benchmarks, the prompt asks the model to select one answer option and to output only the option letter. We do not ask models to provide explanations or chain-of-thought rationales, since the response-side model only requires binary correctness labels. For each item, we evaluate the model under the modality conditions used in Section 3, while keeping the answer options and output format fixed across conditions.

Table 4 summarizes the information exposed to the model under each condition.

The full multimodal prompt is used to measure model performance under the original benchmark setting:

You are solving a multiple-choice question. Output exactly one uppercase letter from [A, B, C, ...] and nothing else.
Mode: image + text.
Use both the provided image(s) and question text.
Question: {question}
Options: {options}
Answer:

The visual-conditioned prompt removes the original question text while retaining the image and answer options:

You are solving a multiple-choice question. Output exactly one uppercase letter from [A, B, C, ...] and nothing else.
Mode: image only.
Question text is hidden. Use only the provided image(s) and options.
Options: {options}
Answer:

The textual-conditioned prompt removes the image while retaining the original question and answer options:

You are solving a multiple-choice question. Output exactly one uppercase letter from [A, B, C, ...] and nothing else.
Mode: text only.
No image is available. Use only question text and options.
Question: {question}
Options: {options}
Answer:

The no-modality prompt removes both the image and the question text while retaining the answer options:

You are solving a multiple-choice question. Output exactly one uppercase letter from [A, B, C, ...] and nothing else.
Mode: none.
No image and no question text are available.
Choose the most plausible answer only from the options below. Guess if needed.

Options: {options}

Answer:

For all prompts, {question} denotes the original benchmark question and {options} denotes the answer options in their original order. We use deterministic decoding when supported by the model API or inference backend. After generation, we extract the first valid option letter from the model output. If no valid option letter can be extracted, the response is treated as incorrect. The extracted answer is compared against the ground-truth answer to obtain the binary response used by the response model.

D.6 Implementation details

We implement SRCoD using Pytorch [27]. For the structural side, we use the LSMI coordinates (R, U_v, U_t, S) . Each coordinate is first clipped at the 0.5-th and 99.5-th percentiles and then standardized across items before being projected into the response-coordinate space. The LSMI-side profile prior is constructed from robustly normalized answerability scores and is used together with the entropy-based confidence weight in the profile-level structural alignment term.

For the response side, we use four latent coordinates corresponding to the base, visual, textual, and joint visual-textual components. In the bounded parameterization, model abilities are constrained to be positive, with the base ability restricted to $[10^{-4}, 0.1]$ and the remaining modality-related abilities restricted to $[10^{-4}, 4]$. Item discriminations are obtained by a softplus transformation and clipped to $[10^{-4}, 4]$. This parameterization keeps the latent response surface numerically stable while preserving the intended interpretation of modality-specific gains.

We optimize the resulting objective with Adam using learning rate 0.01, batch size 2048, and 300 epochs. All experiments use a fixed random seed of 42.

E Case Study

E.1 Redundant Items

Figure 7 presents examples labeled as *Redundant*. In our annotation protocol, a redundant item is one for which both controlled views are sufficient for human annotators to select the correct answer. The visual-conditioned view contains the image and answer options, while the textual-conditioned view contains the original question and answer options. Thus, redundancy should be interpreted as controlled evidence sufficiency rather than semantic equivalence between modalities: the two views may support the same answer through different cues.

ScienceQA. The ScienceQA example asks for the capital of Wyoming, with options Phoenix, Baton Rouge, Honolulu, and Cheyenne. The textual-conditioned view is sufficient because the question explicitly specifies Wyoming, and the options are candidate city names. A human annotator with the relevant factual knowledge can identify Cheyenne as the capital.

The visual-conditioned view is also sufficient because the map highlights Wyoming. Although the original question is not shown in this view, the highlighted state and the city-name options strongly indicate that the item concerns the state shown on the map. Since the options are U.S. state capitals, the highlighted state can be matched to Cheyenne. Therefore, both controlled views provide enough evidence to recover the answer, and the item does not require joint reasoning over image and text.

MMMU. The MMMU example asks what kind of pest is damaging a rose bud, with options including mites, leafhoppers, aphids, and psyllids. The visual-conditioned view provides a direct visual cue: the image shows a rose bud covered with many small clustered insects, a pattern characteristic of aphid infestation. Together with the pest-category options, this is sufficient for selecting aphids.

The textual-conditioned view supports the answer in a different way. The question specifies that the damaged object is a rose bud and restricts the candidates to common plant pests. Among the listed choices, aphids are a highly salient pest associated with rose buds. Thus, the answer can be inferred from the question, options, and domain priors even without relying on the image.

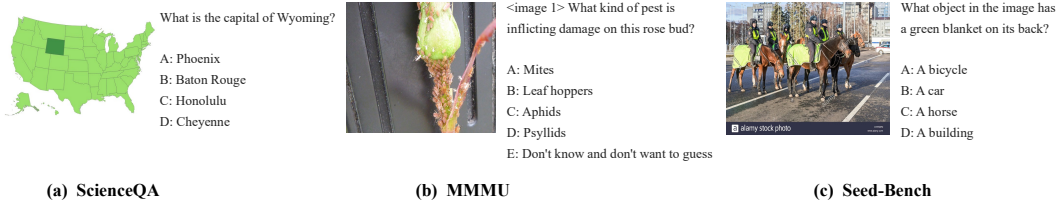


Figure 7: Examples of redundant items from ScienceQA, MMMU, and SEED-Bench.

This example shows why redundant items are not necessarily cases where the two modalities contain the same information. Here, the visual view supports the answer through visible pest appearance, while the textual view supports it through semantic constraints and prior knowledge. Both views are sufficient, but for different reasons.

SEED-Bench. The SEED-Bench example asks which object in the image has a green blanket on its back, with options bicycle, car, horse, and building. The textual-conditioned view is sufficient because the phrase “has a green blanket on its back” strongly constrains the plausible answer: among the options, horse is the only object that naturally has a back and can plausibly carry a blanket.

The visual-conditioned view is sufficient for a different reason. In this view, the annotator does not see the original question, and therefore cannot know that the intended query concerns a green blanket. Nevertheless, the image is dominated by mounted horses, which are the most salient objects among the answer options. Although other option categories such as buildings may also appear in the background, they are not the central objects depicted in the image. Thus, the image-option alignment strongly favors horse even without the original question text.

Summary. This example illustrates a shortcut-like form of redundancy. The item is labeled redundant because both controlled views allow annotators to recover the correct option, but the two views do so through different mechanisms: the textual-conditioned view uses the semantic constraint introduced by the question, whereas the visual-conditioned view relies on salient object-option alignment. Therefore, the item can be answered without genuine cross-modal integration, even though the original full item appears to require matching a textual attribute to a visual object.

E.2 Textual-Dominant Items

Figure 8 presents examples labeled as *Textual-dominant*. In our annotation protocol, a textual-dominant item is one for which the textual-conditioned view is sufficient for human annotators to select the correct answer, while the visual-conditioned view is not. This means that the original question and answer options provide enough semantic, factual, or commonsense constraints to recover the answer, whereas the image-option view alone does not specify the intended query. Textual dominance therefore often reveals cases where the benchmark item can be solved primarily from question wording, domain knowledge, or answer-option priors, rather than from multimodal integration.

ScienceQA. The ScienceQA example asks which organism contains matter that was once part of phytoplankton, with options black rockfish and sea otter. The textual-conditioned view is sufficient because the question explicitly specifies the relevant ecological concept: matter transfer from phytoplankton through a food chain. Given the options, a human annotator can use basic biology knowledge to infer that black rockfish is more directly associated with a phytoplankton-supported aquatic food chain. The text therefore provides a clear reasoning target and enough semantic constraints to select the correct answer.

The visual-conditioned view is less reliable. Although the image shows a food-web diagram containing several organisms and energy-transfer arrows, the original question is not available in this view. With only the image and the options, the annotator does not know which relation in the food web is being queried: the item could ask about predation, energy transfer, trophic level, producer-consumer relations, or matter flow from a particular organism. Because multiple plausible questions are compatible with the same image and options, the visual-conditioned view does not provide a sufficiently

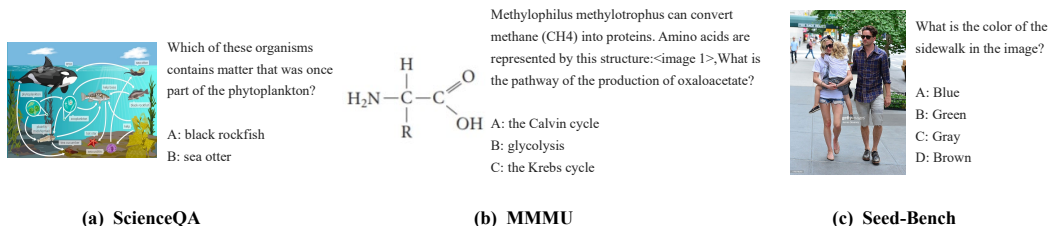


Figure 8: Examples of textual-dominant items from ScienceQA, MMMU, and SEED-Bench.

well-defined decision criterion. This makes the item textual-dominant: the question text determines the relevant ecological relation, while the image alone does not identify what should be answered.

MMMU. The MMMU example asks for the pathway involved in the production of oxaloacetate, with options including the Calvin cycle, glycolysis, and the Krebs cycle. The textual-conditioned view is sufficient because the key term “oxaloacetate” directly invokes biochemical knowledge. Oxaloacetate is a central intermediate in the Krebs cycle, so the correct answer can be inferred from the question and options without relying on the image. The surrounding context about methane conversion and amino acids further frames the question as a biochemical pathway problem.

In contrast, the visual-conditioned view provides only a generic amino acid structure and the answer options. This structure indicates the general form of amino acids, but it does not specify oxaloacetate or the metabolic pathway being queried. Without the textual cue, the options Calvin cycle, glycolysis, and Krebs cycle are all plausible biochemical processes, and the image does not determine which pathway should be selected. Thus, the answer is driven primarily by textual biochemical knowledge rather than by visual evidence.

SEED-Bench. The SEED-Bench example asks for the color of the sidewalk in the image, with options blue, green, gray, and brown. The textual-conditioned view is sufficient because the question explicitly names the target object, “sidewalk”, and the answer options are colors. Even without the image, common visual priors make gray the most plausible answer, since sidewalks are typically gray. The text and options therefore create a strong answer prior.

The visual-conditioned view, however, is ambiguous. The image contains several visible colors: green vegetation, gray pavement, blue clothing, and other background colors. Because the original question is absent, the annotator does not know which object or region should be inspected. The correct option is not determined by the image-option view alone; it becomes clear only when the text specifies that the target is the sidewalk. This item is therefore textual-dominant because the decisive cue is the textual specification of the queried object, while the image without the question leaves the selection underdetermined.

Summary. These examples show that textual-dominant items may appear multimodal in format, but their answers can be recovered primarily from the question and answer options. Such items are not necessarily invalid for all purposes, but they provide limited evidence that a model must use visual information. They are especially important for benchmark diagnosis because they can inflate multimodal evaluation scores through textual shortcuts, domain priors, or option-driven reasoning.

E.3 Visual-Dominant Items

Figure 9 presents examples labeled as *Visual-dominant*. In our annotation protocol, a visual-dominant item is one for which the visual-conditioned view is sufficient for human annotators to select the correct answer, while the textual-conditioned view is not. The visual-conditioned view contains the image and answer options, whereas the textual-conditioned view contains the original question and answer options. Thus, visual dominance indicates that the decisive evidence comes from the image-option view: without the image, the question either lacks the referenced object, lacks the visual trend, or lacks the visual attribute needed to identify the answer.

ScienceQA. The ScienceQA example asks which property matches the object, with options *colorful* and *sticky*. The visual-conditioned view is sufficient because the image clearly depicts a kite with

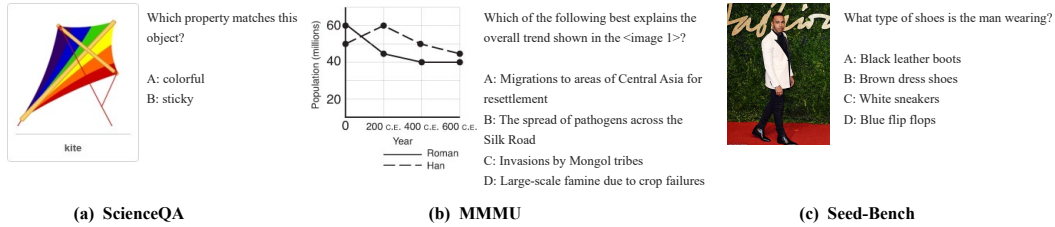


Figure 9: Examples of visual-dominant items from ScienceQA, MMMU, and SEED-Bench.

multiple bright colors. Given the answer options, the visible appearance of the object directly supports *colorful* as the correct choice. The image also contains the object label “kite”, which further makes the visual view self-contained for this simple property-matching question.

The textual-conditioned view is insufficient because the question only refers to “this object” without showing the object. From the text and options alone, annotators know that they must choose a property, but they do not know what object is being described. Both options are generic properties that could apply to different objects in different contexts. Therefore, the correct answer cannot be reliably determined without the visual input. This item is visual-dominant because the image identifies the object and its salient property, while the text alone does not specify the target.

MMMU. The MMMU example asks which option best explains the overall trend shown in a population chart. The visual-conditioned view is sufficient because the chart shows the population trajectories of the Roman and Han populations over time. Both series exhibit an overall decline, and the answer options provide candidate historical explanations. By inspecting the plotted trend together with the options, annotators can infer that the question is asking for an explanation of the broad decline and select the option most consistent with that historical pattern.

The textual-conditioned view is not sufficient because the question depends on the phrase “the overall trend shown in the image”. Without the chart, annotators cannot observe what the trend is: it could be an increase, a decline, a divergence between the two populations, or some other pattern. Although the options mention possible historical causes, the text alone does not specify which empirical pattern needs to be explained. As a result, the answer cannot be grounded from the question and options alone. This item is therefore visual-dominant because the decisive information is the visual trend in the chart.

SEED-Bench. The SEED-Bench example asks what type of shoes the man is wearing, with options including black leather boots, brown dress shoes, white sneakers, and blue flip flops. The visual-conditioned view is sufficient because the image clearly shows the man and his footwear. Given the options, the shoes can be identified from the visual appearance as black leather boots. The answer is determined by a local visual attribute of the person in the image.

The textual-conditioned view is insufficient because the question asks about an attribute of “the man” but provides no visual evidence about his appearance. The answer options list plausible footwear types, and none can be selected reliably from the wording alone. Any choice would amount to guessing or relying on an unsupported prior. Thus, the item is visual-dominant: the image provides the necessary attribute evidence, while the text only specifies what kind of visual attribute should be inspected.

Summary. These examples illustrate a common form of visual dominance in multimodal benchmarks. The question text may define the type of query, but the answer is determined by visual evidence such as object appearance, chart structure, or local attributes in the image. Without the image, the item is under-specified; with the image and answer options, the correct answer becomes identifiable.

E.4 Synergistic Items

Figure 10 presents examples labeled as *Synergistic*. In our annotation protocol, a synergistic item is one for which neither the visual-conditioned view nor the textual-conditioned view is sufficient

for human annotators to select the correct answer, while the original full multimodal item remains solvable. Intuitively, synergistic items require combining the semantic target specified by the question with the visual evidence provided by the image. The text tells the annotator what relation, attribute, or concept should be identified, while the image provides the concrete objects, structures, or values needed to ground that query. When either side is removed, the answer becomes underdetermined.

ScienceQA. The ScienceQA example asks which organism is the tertiary consumer in a food web. The textual-conditioned view is insufficient because, although the question specifies the ecological concept of interest, it does not provide the food-web structure needed to determine which organism occupies the relevant trophic role. From the options alone, an annotator cannot infer which organism is connected to which others or how energy flows through the system.

The visual-conditioned view is also insufficient. The food-web diagram contains multiple organisms and feeding relations, but without the original question, the annotator does not know what aspect of the diagram should be identified. The same image could support many different questions, such as asking for a producer, a primary consumer, a predator, or an organism that eats a specific prey. Thus, the image and options alone do not specify that the target concept is “tertiary consumer.”

The item becomes answerable only when the two modalities are combined. The question contributes the reasoning target—the trophic role to be identified—and the image provides the food-web structure needed to trace the corresponding organism. This is a characteristic synergistic case: the answer depends on combining conceptual information from the text with relational information from the image.

MMMU. The MMMU example asks which numbered position on an action-potential graph represents the resting potential. The textual-conditioned view is insufficient because the concept “resting potential” is defined relative to the visual graph and its numbered markers. Without the graph, the annotator cannot determine which numbered positions correspond to baseline membrane potential, depolarization, repolarization, or hyperpolarization.

The visual-conditioned view is likewise insufficient. The graph shows a membrane-potential curve with numbered markers, but without the original question, the annotator does not know which feature of the curve is being queried. The item could ask about resting potential, threshold, peak depolarization, repolarization, or after-hyperpolarization. Although the graph contains all of this information, the intended target cannot be identified from the image-option view alone.

Only the joint multimodal view resolves the task. The text specifies the electrophysiological concept to locate, while the image provides the graph and numbered positions that instantiate that concept. Neither modality alone is enough, but their combination yields a well-defined and answerable question. This is precisely the pattern captured by the synergistic label.

SEED-Bench. The SEED-Bench example asks for the color of the chair next to the blue couch. The textual-conditioned view is insufficient because the question specifies a target object “the chair next to the blue couch” relationally, but without the image, the annotator cannot see the room layout or identify the color of that chair. Although the answer options list candidate colors, the text alone does not provide enough information to determine which one is correct.

The visual-conditioned view is also insufficient. The room image contains several colored objects, including a blue couch, a purple chair, yellow accents, and other furniture. However, without the original question, the annotator does not know which object attribute is being queried. The image alone does not indicate whether the task concerns the couch, the chair, the walls, or some other object, nor does it specify that the relevant chair is the one adjacent to the blue couch.

The full multimodal item resolves this ambiguity. The text identifies the target through a relational description, and the image grounds that description by showing which chair is next to the blue couch and what color it is. This item is therefore synergistic: the answer depends on combining the textual reference with the visual scene.

Summary. These examples illustrate the core diagnostic meaning of synergy in multimodal benchmarks. A synergistic item is not merely one that contains both image and text, but one in which the two modalities make complementary contributions that are both necessary for answering the

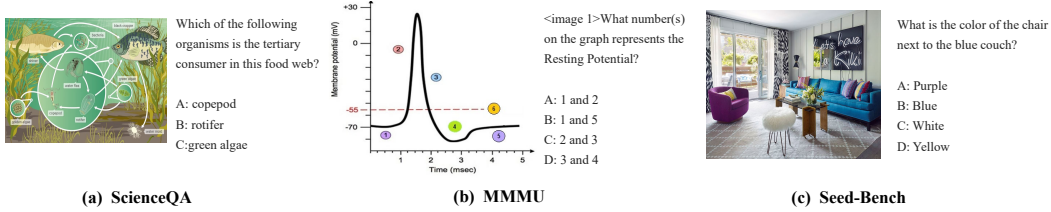


Figure 10: Examples of synergistic items from ScienceQA, MMMU, and SEED-Bench.

Table 5: Paired statistical significance analysis of SRCoD over M3IRT on the human-labeled diagnostic validation set. We report paired performance differences, 95% bootstrap confidence intervals, approximate-randomization p -values, and McNemar’s test for accuracy.

Dataset	Accuracy			Macro-F1			McNemar
	Δ	95% CI	p_{rand}	Δ	95% CI	p_{rand}	
ScienceQA	+0.037	[0.015, 0.060]	< 0.001	+0.049	[0.026, 0.075]	< 0.001	87/50, $p = 0.002$
MMMU	+0.061	[0.035, 0.089]	< 0.001	+0.054	[0.025, 0.079]	< 0.001	129/68, $p < 0.001$
SEED-Bench	+0.067	[0.042, 0.091]	< 0.001	+0.070	[0.043, 0.095]	< 0.001	115/48, $p < 0.001$
Average	+0.055	[0.040, 0.069]	< 0.001	+0.058	[0.044, 0.072]	< 0.001	331/166, $p < 0.001$

question. Such items provide the strongest evidence that a benchmark is actually testing multimodal understanding rather than unimodal shortcuts or answer-option priors.

F Experiment Extension

F.1 Statistical Significance Analysis

We further conduct paired statistical significance analysis on the human-labeled diagnostic validation set. Since all diagnosis methods are evaluated on the same set of items, we compare SRCoD with the strongest baseline, M3IRT, using paired item-level tests. For both accuracy and macro-F1, we report the paired performance difference, 95% confidence intervals estimated by item-level bootstrap resampling, and paired approximate-randomization p -values. For accuracy, we additionally report McNemar’s test, which directly compares the number of items correctly classified only by SRCoD and only by M3IRT.

Table 5 shows that the improvements of SRCoD over M3IRT are statistically significant across all three benchmarks. On ScienceQA, SRCoD improves accuracy by +0.037 with a 95% confidence interval of [0.015, 0.060], and improves macro-F1 by +0.049 with a 95% confidence interval of [0.026, 0.075]. On MMMU and SEED-Bench, the gains are larger: SRCoD improves accuracy by +0.061 and +0.067, respectively, and improves macro-F1 by +0.054 and +0.070, respectively. All bootstrap confidence intervals are strictly positive, and all approximate-randomization tests yield $p < 0.001$.

The McNemar results provide a complementary paired comparison for accuracy. For each benchmark, the number of items correctly classified only by SRCoD is substantially larger than the number correctly classified only by M3IRT. For example, on MMMU, SRCoD uniquely corrects 129 items, compared with 68 items uniquely corrected by M3IRT; on SEED-Bench, the corresponding counts are 115 and 48. The McNemar p -values are below 0.01 for all benchmarks. Aggregating across the validation items, SRCoD improves accuracy by +0.055 and macro-F1 by +0.058, with both confidence intervals remaining strictly positive and all paired significance tests yielding $p < 0.001$. These results indicate that the observed gains of SRCoD are unlikely to be explained by item-level sampling variation.

F.2 Comparison with Simple Fusion

To examine whether the improvement of SRCoD can be attributed to a simple combination of intrinsic structural information and extrinsic model responses, we compare SRCoD with a post-hoc profile fusion baseline. Unlike SRCoD, which jointly calibrates structural priors and response-based item

Table 6: Comparison with post-hoc profile fusion baselines. We report accuracy and macro-F1 on the human-labeled validation set.

Method	ScienceQA		MMMU		SEED-Bench		Average	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
$\alpha = 0.25$	0.52	0.44	0.46	0.46	0.61	0.60	0.53	0.50
$\alpha = 0.50$	0.47	0.36	0.45	0.44	0.60	0.59	0.51	0.46
$\alpha = 0.75$	0.44	0.30	0.45	0.44	0.60	0.58	0.50	0.44
SRCoD	0.61	0.59	0.55	0.54	0.69	0.67	0.61	0.60

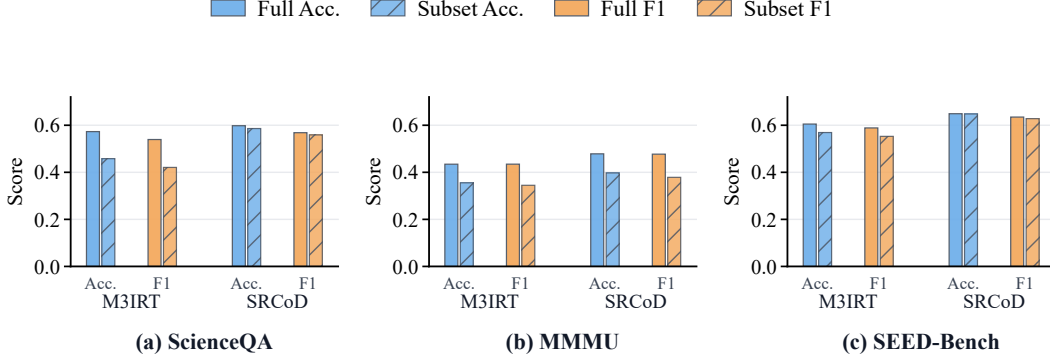


Figure 11: Diagnostic performance under a small-scale model pool. We compare M3IRT and SRCoD using responses from 3B-12B MLLMs and report accuracy and macro-F1. SRCoD consistently outperforms M3IRT and shows smaller degradation on the subset, indicating stronger robustness when response-side evidence is obtained from smaller and noisier models.

parameters, this baseline independently obtains the two diagnostic profiles and combines them using a fixed global weight.

Specifically, the fused diagnostic profile for item j is defined as

$$\mathbf{q}_j^{\text{fuse}} = \alpha \mathbf{q}_j^{\text{str}} + (1 - \alpha) \mathbf{q}_j^{\text{resp}}, \quad (28)$$

where $\mathbf{q}_j^{\text{str}}$ is the normalized four-factor structural profile derived from LSMI, and $\mathbf{q}_j^{\text{resp}}$ is the normalized response-based profile produced by M3IRT. The coefficient α controls the relative contribution of structural information. We evaluate three representative settings, $\alpha \in \{0.25, 0.50, 0.75\}$, and determine the dominant diagnostic factor from the resulting fused profile.

Table 6 shows that SRCoD consistently outperforms all three fusion settings across the benchmarks. Even the strongest fusion baseline, with $\alpha = 0.25$, achieves an average accuracy of 0.53 and macro-F1 of 0.50, compared with 0.61 and 0.60 for SRCoD, respectively.

These results indicate that the improvement of SRCoD cannot be reproduced by the tested fixed-weight fusion strategies. By incorporating structural information into item-parameter estimation and profile-level calibration, SRCoD allows intrinsic structural evidence to interact with observed model responses during optimization. This provides a more effective integration of the two diagnostic signals than combining independently estimated profiles.

F.3 Robustness under a Small-Scale Model Pool

To evaluate whether SRCoD remains reliable when response-side evidence is limited to smaller models, we construct a small-scale MLLM pool with comparable parameter sizes, ranging from 3B to 12B parameters. This setting is more challenging than using a diverse pool of strong MLLMs, since smaller models often produce noisier and more biased modality-conditioned responses. We fit both M3IRT and SRCoD using responses from this small-scale model pool and evaluate their diagnostic performance on the human-labeled validation set.

Figure 11 shows that SRCoD remains consistently stronger than M3IRT when both methods are fitted using only responses from the small-scale model pool. On the full validation set, SRCoD outperforms M3IRT across ScienceQA, MMMU, and SEED-Bench in both accuracy and macro-F1. The advantage is more pronounced on the subset evaluation. For example, on ScienceQA, M3IRT shows a substantial drop from the full set to the subset, whereas SRCoD maintains nearly the same performance. A similar pattern appears on SEED-Bench, where SRCoD remains stable while M3IRT degrades. On MMMU, both methods drop on the subset, but SRCoD still achieves higher accuracy and macro-F1.

These results suggest that SRCoD is less dependent on the reliability of response-side evidence alone. Since M3IRT relies purely on model response patterns, its estimates become less stable when the model pool consists of smaller and noisier MLLMs. SRCoD mitigates this issue by anchoring item parameters with intrinsic item structure, leading to more robust item-level diagnosis under limited-model settings.

F.4 Behavioral Validation on Held-Out Model Families

The model-pool subsampling experiment in Section 4.4 examines the stability of SRCoD under reduced diagnostic model pools. We further investigate whether the predicted modality-dependence profiles exhibit their expected behavioral patterns on model families that are not involved in SRCoD fitting.

Experimental Protocol. We conduct a leave-one-model-family-out evaluation on ScienceQA and SEED-Bench, using Qwen and Ministral as two held-out model families. For each benchmark, we construct two family-level holdout settings. In the Qwen setting, all responses from Qwen-family models are excluded from SRCoD fitting. Similarly, all Ministral-family responses are excluded in the Ministral setting.

SRCoD is fitted exclusively using responses from the remaining model families. Each item is then assigned to the diagnostic category corresponding to its highest calibrated profile score. Importantly, responses from the held-out family are not used for either model fitting or item-profile assignment. They are used only for subsequent behavioral evaluation.

Behavioral Evaluation. We evaluate each held-out model family under four modality conditions: no modality (0), image only (v), text only (t), and image-text (vt). For each item and condition, correctness is first averaged across models within the held-out family. We then aggregate these item-level accuracies within each SRCoD-predicted diagnostic category.

For a given predicted category, let A_c denote the mean held-out-family accuracy under modality condition c . We define four behavioral margins that characterize the expected response pattern of each diagnostic category:

$$\begin{aligned} G_{\text{vis}} &= A_v - \max(A_t, A_0), \\ G_{\text{text}} &= A_t - \max(A_v, A_0), \\ G_{\text{red}} &= \min(A_v - A_0, A_t - A_0), \\ G_{\text{syn}} &= A_{vt} - \max(A_v, A_t). \end{aligned} \tag{29}$$

Each margin is computed on the item group assigned to its corresponding diagnostic category. A positive margin indicates that the held-out family exhibits the expected behavioral direction. In particular, $G_{\text{vis}} > 0$ and $G_{\text{text}} > 0$ indicate that the corresponding unimodal condition outperforms the alternatives. Similarly, $G_{\text{red}} > 0$ requires both unimodal conditions to outperform the no-modality baseline, while $G_{\text{syn}} > 0$ indicates an advantage of joint image-text input over either unimodal condition.

Results. Table 7 shows that all 16 behavioral margins are positive across the two held-out model families and two benchmarks. In particular, items diagnosed as Synergistic consistently exhibit higher average accuracy under joint image-text input than under either unimodal condition. The Visual-dominant, Textual-dominant, and Redundant groups likewise exhibit their expected modality-specific behavioral patterns.

These results indicate that the diagnostic categories predicted by SRCoD preserve their expected aggregate behavioral characteristics on unseen model families. This provides further evidence that

Table 7: Behavioral validation on held-out model families. Values are absolute accuracy margins computed within the corresponding SRCoD-predicted diagnostic groups. Positive values indicate agreement with the expected behavioral patterns.

Held-out Family	Benchmark	G_{vis}	G_{text}	G_{red}	G_{syn}
Qwen	SEED-Bench	+0.28	+0.11	+0.26	+0.15
Qwen	ScienceQA	+0.37	+0.40	+0.37	+0.11
Ministral	SEED-Bench	+0.44	+0.20	+0.18	+0.23
Ministral	ScienceQA	+0.30	+0.35	+0.27	+0.13

structure-response co-calibration captures modality-dependence patterns that generalize beyond the diagnostic model pool used during fitting.

G Broader Impacts

This work aims to improve the reliability and interpretability of multimodal benchmark evaluation. By diagnosing whether individual benchmark items rely on redundant, visual-dominant, textual-dominant, or synergistic evidence factors, the proposed method can help benchmark developers identify shortcut-prone or weak items and construct more informative evaluation sets. This may contribute to more faithful assessment of vision-language models and reduce the risk of overestimating cross-modal reasoning ability from aggregate benchmark scores.

The work also has potential negative impacts if used improperly. Diagnostic results may be affected by the choice of model pool, the quality of intrinsic structural estimates, and uncertainty in human evidence-sufficiency annotations. If such diagnostics are treated as definitive rather than as supporting evidence, benchmark developers may incorrectly discard useful items or overemphasize particular types of modality dependence. In addition, improving benchmark filtering may indirectly accelerate the development of stronger multimodal systems, which could be misused in downstream applications involving misinformation, surveillance, or biased decision support. To mitigate these risks, we recommend using the proposed diagnosis as an auxiliary auditing tool, reporting uncertainty where possible, and combining automatic diagnosis with human review before making high-stakes benchmark or deployment decisions.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims in the abstract and introduction accurately reflect the paper’s contributions and scope. The main methodological claims are supported by Section 3, and the empirical claims are supported by the results in Section 4.

Guidelines:

- The answer [\[N/A\]](#) means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [\[No\]](#) or [\[N/A\]](#) answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of our work in Appendix A.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include formal theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclose the information needed to reproduce the main experimental results in Section 4 and also provide an anonymized repository linked in the abstract, which contains the code and relevant datasets.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide an anonymized repository linked in the abstract, which contains the code and relevant datasets.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: The training and test details are specified in Section 4.1, including the experimental setup, data splits, hyperparameters, and implementation details necessary to understand the reported results.

Guidelines:

- The answer [\[N/A\]](#) means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: The paper reports statistical significance analyses in Appendix F.1.

Guidelines:

- The answer [\[N/A\]](#) means that the paper does not include experiments.
- The authors should answer [\[Yes\]](#) if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.).
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: The compute resources used for the experiments are described in Section 4.1 and Appendix, including the GPU type and relevant computational setup used to run the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and confirm that the research conducted in this paper conforms to it.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix G discusses both positive and negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data or models that pose a high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use existing public datasets and properly credit their original creators in the references. The licenses and terms of use for these datasets are explicitly reported in Appendix, and our use complies with those terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a new human-labeled diagnostic validation set for item-level multimodal benchmark diagnosis. The dataset will be released through an anonymized GitHub link.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[Yes\]](#)

Justification: The paper involves crowdsourced human annotation for constructing the diagnostic validation set. Appendix C provides the annotation protocol. Annotators were asked to answer multiple-choice benchmark items under controlled modality conditions, and no personal or sensitive information was collected.

Guidelines:

- The answer [\[N/A\]](#) means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[No\]](#)

Justification: The paper involves paid human annotation for constructing the diagnostic validation set, but no formal IRB or equivalent institutional review was obtained. The annotation task is minimal-risk: participants only answer multiple-choice benchmark items under controlled modality conditions, and no personal, sensitive, or health-related information is collected.

Guidelines:

- The answer [\[N/A\]](#) means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [\[Yes\]](#)

Justification: MLLMs are used as diagnostic respondents to collect modality-conditioned response matrices for fitting and evaluating the diagnosis methods. The model pool, prompts, modality conditions, and response normalization procedures are documented in Appendix D. LLMs were additionally used only for language polishing, which did not affect the methodology or scientific claims.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.