SPAR: LEARNING VLM RISK FROM STRUCTURED PEER AGREEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models can be confidently wrong, and their confidence scales transfer poorly across architectures and tasks. We study *fixed-focal selective prediction*, where heterogeneous peer outputs help decide whether to accept or abstain on a focal VLM answer without changing its content. Common multi-VLM uncertainty scores compress this evidence into scalar agreement, even though the same mean disagreement can arise when peers agree with one another but oppose the focal answer, or when they disagree among themselves. We introduce SPAR, a **S**tructured **P**eer-**A**greement-based **R**isk estimator for fixed-focal selective prediction. It decomposes focal disagreement into focal isolation and peer ambiguity, and combines this decomposition with relative peer confidence. Percentile-rank alignment makes continuous features comparable across models and datasets, allowing a single risk scorer to be applied unchanged across different settings. Across five focal VLMs and three multimodal benchmarks, SPAR achieves the best aggregate AURC and consistently outperforms the evaluated peer-based and repeated-sampling alternatives. Under simultaneous model-family and dataset shift, it reduces macro-AURC by 7.05% over the strongest baseline (Consensus Entropy), with gains in every setting, while achieving up to a 10.1× generation-time speedup over the strongest repeated-sampling comparison. These results establish structured output relations as an effective and transferable reliability signal for selective VLM prediction.

1 INTRODUCTION

Vision–language models (VLMs) can produce fluent, plausible answers even when they misread the image or fail to ground the question. In high-stakes domains such as medicine, law, and transportation, an incorrect response can misdirect consequential decisions, making it essential to identify risky answers and abstain. Selective prediction provides this safeguard by ranking answers according to failure risk and abstaining when that risk is high (Geifman & El-Yaniv, 2017; Whitehead et al., 2022; Wieczorek et al., 2026; Sarkar et al., 2026). We study *fixed-focal selective prediction*, where the focal VLM has already produced the deployed answer and a risk scorer assigns it a failure-risk score for acceptance or abstention, without revising or replacing the answer. This interface cleanly separates risk estimation from answer generation. To transfer beyond its training setting, the risk scorer must remain effective when the focal model family and dataset both change.

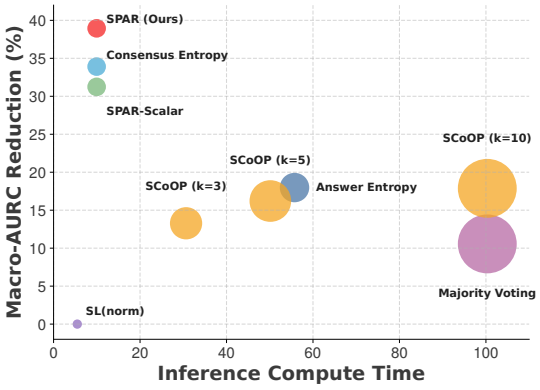


Figure 1: The x-axis shows total VLM inference time per example (seconds), and the y-axis shows macro-AURC reduction relative to SL(norm). Marker size denotes the total number of VLM inference calls per example.

Existing failure detectors estimate reliability from response consistency across proxy-generated question rephrasings (Khan & Fu, 2024), learned representations (Lafon et al., 2025; Kogilathota

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2026), evidential features (Huang et al., 2026), or attention statistics (Hoang et al., 2026). These methods provide useful uncertainty signals, but often require repeated queries, model-specific internal access, or an auxiliary learned model. Multi-model methods instead exploit independent VLM judgments for verification, abstention, answer selection, or routing (Sun et al., 2024; Zhang et al., 2026b; Yu et al., 2026; Zhang et al., 2026a). Aggregation and routing, however, can change the prediction being evaluated. When the focal answer must remain fixed, a common approach is to compress peer evidence into a scalar consensus score (Zhang et al., 2026b; Yu et al., 2026). This leaves a fundamental question: can a single query to each of a few heterogeneous VLMs provide a unified and transferable risk signal for a focal answer?

The key limitation of mean focal–peer disagreement is that it only captures magnitude but not structure. It measures the average disagreement between the focal answer and the peer predictions but does not describe how the peer predictions relate to one another. The same mean focal–peer disagreement can reflect either coherent peer opposition to the focal answer or inconsistency within the peer models. Although indistinguishable to a scalar score, these structures may provide different evidence of focal error. Recent semantic uncertainty estimators likewise preserve structure among sampled answers using semantic-similarity kernels, weighted graphs, or hierarchical encoding trees (Nikitin et al., 2024; Li et al., 2025; Zhao et al., 2025). Related LVLM work further shows that varied but coherent responses and directly contradictory responses represent different kinds of uncertainty (Shi et al., 2026). Thus, estimating focal error should use both focal–peer disagreement and peer–peer agreement instead of compressing the peer evidence into a single scalar.

We introduce SPAR, a framework that learns VLM risk through Structured Peer Agreement-based Risk estimation. SPAR constructs an answer-relation graph over focal–peer and peer–peer edges, and exactly decomposes mean focal disagreement into *focal isolation* and *peer ambiguity*. It combines these relations with relative peer confidence, aligns model-specific features using percentile ranks computed from unlabeled examples, and maps the resulting representation to risk with a scorer trained on source tasks. Together, these operations preserve information suppressed by scalar consensus while making heterogeneous confidence scales comparable. SPAR requires only one deterministic answer and its generated-answer likelihood from each model; it uses no hidden activations, gradients, VLM fine-tuning, auxiliary judge, target correctness labels, or answer replacement. SPAR reduces macro-AURC by 7.05% under simultaneous model-family and dataset shift, with gains in every target setting. Figure 1 further shows that SPAR combines strong risk ordering with low inference cost, achieving a $10.1\times$ generation-time speedup over SCoOP at its largest sampling budget.

Our main contributions are summarized as follows:

- We show that scalar summaries of multi-VLM consensus obscure whether disagreement reflects focal isolation or peer ambiguity. We separate these cases through an exact decomposition of mean focal disagreement, then incorporate relative peer confidence and percentile-rank alignment across heterogeneous models.
- We introduce SPAR, an efficient fixed-focal reliability framework that learns selective risk from a confidence-aware representation built from an answer-relation graph, requiring one response per model and leaving the deployed answer unchanged.
- We validate SPAR across model families, datasets, and inference budgets, including simultaneous model-family and dataset shift and focal VLMs spanning 4B–90B parameters. Controlled comparisons isolate the value of structured relations, while deployment analyses demonstrate robustness across rank references and peer budgets.

2 RELATED WORK

Selective prediction and VLM failure detection. Selective prediction equips a predictor with a reject option and evaluates confidence through the trade-off between the error rate on accepted answers and the fraction of answers accepted (Geifman & El-Yaniv, 2017). In VQA, learned selectors estimate answer reliability from the deployed model or from peer predictions available during training (Whitehead et al., 2022; Dancette et al., 2023). Other methods use variational selective learning or retrieval-based memory to estimate answer reliability (Wieczorek et al., 2026; Sarkar et al., 2026). A complementary line of VLM failure detectors draws on black-box consistency (Khan & Fu, 2024), learned vision–text representations or embedding geometry (Lafon et al., 2025; Dong et al., 2025),

Bayesian posteriors (Baumann et al., 2026), or internal evidence and attention (Kogilathota et al., 2026; Huang et al., 2026; Hoang et al., 2026). These methods require different forms of model access, including repeated queries, learned embeddings, and model-internal states. By contrast, SPAR estimates the risk of a fixed focal answer using only the outputs and generated-answer likelihoods of frozen heterogeneous VLMs.

Sampling-based and structured semantic uncertainty. Black-box uncertainty commonly measures consistency across repeated responses (Manakul et al., 2023; Lin et al., 2024b). Semantic Entropy groups meaning-equivalent generations (Kuhn et al., 2023; Farquhar et al., 2024), while Kernel Language Entropy, Semantic Graph Density, and SeSE preserve graded pairwise or hierarchical relations (Nikitin et al., 2024; Li et al., 2025; Zhao et al., 2025). Recent analyses further distinguish semantic divergence from logical conflict and ambiguity from missing knowledge (Shi et al., 2026; Nan et al., 2025). SPAR similarly preserves relations among responses rather than reducing them to a single consistency score. Instead of comparing repeated generations from one model, it uses one deterministic answer from each heterogeneous VLM and anchors the graph on the focal answer whose risk is being estimated.

Multi-VLM consensus and structured relations. CrossCheckGPT uses agreement between outputs from different models for reference-free hallucination ranking (Sun et al., 2024); Consensus Entropy and SCoOP use multi-VLM agreement for verification, selection, or measuring uncertainty across models (Zhang et al., 2026b; Yu et al., 2026). Multi-agent systems further exploit structured uncertainty or model diversity to route inputs between models and combine their predictions (Zhang et al., 2026a; Tekin et al., 2026). These objectives range from ranking an existing output to constructing a new system answer. In contrast, SPAR operates in the fixed-focal setting: each peer contributes one deterministic output solely to refine the risk ordering of an unchanged focal answer. Its confidence-aware representation built from the answer-relation graph, together with percentile-rank alignment, allows a scorer trained on source model–dataset pairs to be applied to new model families and datasets without refitting.

3 METHOD

3.1 FIXED-FOCAL SELECTIVE PREDICTION

For an input $x_i = (v_i, q_i)$, a focal VLM M_0 produces the answer a_i^0, and K heterogeneous peers $M_1, \ldots, M_K$ produce answers $a_i^1, \ldots, a_i^K$. SPAR assigns the focal answer a risk score $\hat{s}_i$: the answer is accepted below a chosen threshold, and the system abstains otherwise, but its content never changes. A scorer trained on source domains is frozen for all target model–dataset domains.

As illustrated in Figure 2, each frozen VLM is queried once and provides only its answer and likelihood. SPAR neither aggregates peer outputs into a new answer nor requires hidden activations, gradients, model updates, an auxiliary judge, or target-side coefficient fitting. This interface separates reliability estimation from focal-answer generation.

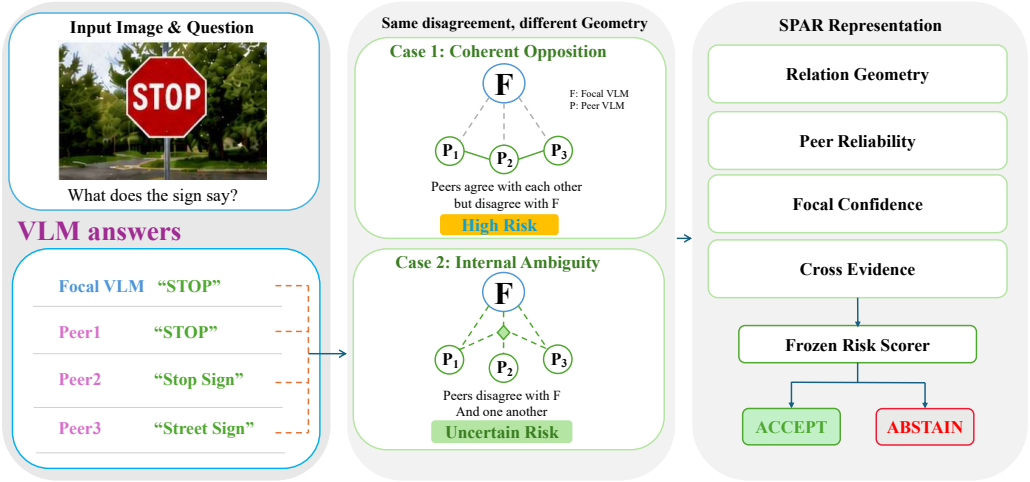
We compute focal–peer and peer–peer similarities from the answers and combine them with confidence signals derived from token-level likelihoods to construct the representation described below.

3.2 ANSWER-RELATION DECOMPOSITION

SPAR forms a complete weighted graph $\mathcal{G}_i = (V_i, E_i)$ with one vertex for the focal answer and one for each peer answer. We compare each pair of answers using token-level F1, following the standard overlap metric used in extractive QA (Rajpurkar et al., 2016). The resulting score $s(a, b) \in [0, 1]$ serves as the edge weight. For $K \geq 2$, we summarize the graph by the mean focal–peer similarity $\bar{s}_{f,i}$ and mean peer–peer similarity $\bar{s}_{p,i}$:

$$\bar{s}_{f,i} = \frac{1}{K} \sum_{k=1}^{K} s(a_i^0, a_i^k), \qquad \bar{s}_{p,i} = \frac{2}{K(K-1)} \sum_{1 \leq k < \ell \leq K} s(a_i^k, a_i^\ell). \tag{1}$$

We define D_i as mean focal disagreement and A_i as peer ambiguity, which measures disagreement among peers. Focal isolation I_i compares agreement among peers with their agreement with the



(a) Fixed Predictions from Heterogeneous VLMs **(b) Structured Relations** **(c) Selective Risk Prediction**

Figure 2: **Overview of SPAR.** Given a fixed focal prediction and heterogeneous peer predictions, SPAR constructs answer relations and exactly decomposes mean focal–peer disagreement D into focal isolation I and peer ambiguity A. Its rank-aligned representation comprises focal confidence, answer-relation geometry, peer-reliability proxies, and cross-evidence interactions. A source-trained scorer maps this representation to selective risk and remains frozen during evaluation, without changing the focal answer.

focal answer. These quantities satisfy:

$$D_i = 1 - \bar{s}_{f,i}, \qquad A_i = 1 - \bar{s}_{p,i}, \qquad I_i = D_i - A_i = \bar{s}_{p,i} - \bar{s}_{f,i}. \tag{2}$$

Peer ambiguity A_i increases as mean peer–peer similarity decreases, while focal isolation I_i measures the excess of peer–peer similarity over focal–peer similarity. At fixed D_i, low A_i and high I_i characterize coherent peer opposition, whereas high A_i and low I_i characterize internal peer disagreement. Beyond these quantities, SPAR further includes summary statistics of focal–peer and peer–peer similarities, exact-match support and opposition, peer centrality, and four focal-relative answer patterns: unanimous support, unanimous opposition, partial support, and fragmented opposition. Complete definitions appear in Section D.

3.3 RANK-ALIGNED RELATIONS

For the answer generated by M_k, let ℓ_{it}^k be the conditional log-probability of its t-th generated answer token, and let T_i^k be the answer length in tokens. We convert the average token log-probability into an answer-level likelihood risk, with larger values indicating lower likelihood:

$$u_i^k = 1 - \exp\left(\frac{1}{T_i^k}\sum_{t=1}^{T_i^k} \ell_{it}^k\right). \tag{3}$$

Likelihood-derived risks and graph-level features can have different scales across models and datasets. We therefore map each feature value to its percentile rank across examples. For a feature z, such as the model-specific risk u^k or a graph-level quantity D, I, or A, observed over an unlabeled reference group G, we define

$$R_G(z_i) = \frac{\mathrm{rank}_{\mathrm{avg}}(z_i; \{z_j : j \in G\}) - \frac{1}{2}}{|G|}. \tag{4}$$

The operator $\mathrm{rank}_{\mathrm{avg}}$ assigns average ranks to tied values. Each graph-level feature is ranked separately across all examples in the current model–dataset setting, denoted by G. For likelihood-derived

risk, we apply the same transformation separately to each model, obtaining $R_{G_k}(u_i^k)$, where G_k contains the reference examples scored by M_k, with G_0 corresponding to the focal model. Because R_G is invariant to strictly increasing reparameterizations, it provides comparable relative features across heterogeneous focal and peer families without target labels.

For each peer $k \geq 1$, we define the rank-aligned confidence proxy $c_i^k = 1 - R_{G_k}(u_i^k)$ and per-peer focal disagreement $d_i^k = 1 - s(a_i^0, a_i^k)$. Their confidence-weighted average is

$$D_{w,i} = \frac{\sum_{k=1}^{K} c_i^k d_i^k}{\sum_{k=1}^{K} c_i^k}. \tag{5}$$

Using the pairwise answer similarities and likelihood-derived risks defined above, SPAR constructs three groups of base features. The focal-confidence component $\phi_i^{\text{focal}} = R_{G_0}(u_i^0)$ captures the focal answer's relative likelihood risk. The answer-relation group ϕ_i^{relation} summarizes focal isolation, peer ambiguity, and mean focal disagreement, together with similarity statistics, support and opposition, peer centrality, and focal-relative answer patterns. The peer-reliability group ϕ_i^{peer} summarizes peer risks and confidence-conditioned relations, including $D_{w,i}$. Continuous base features are rank-aligned, while answer-pattern indicators remain binary.

The cross-evidence group ϕ_i^{cross} is then formed from products of selected aligned features, combining focal risk, answer relations, peer confidence, and peer consistency. Representative interactions are $R_{G_0}(u_i^0)R_G(I_i)$ and $R_G(D_{w,i})[1 - R_G(A_i)]$, where $1 - R_G(A_i)$ measures peer consistency. The four groups are concatenated as

$$\phi_i^{\text{SPAR}} = \left[\phi_i^{\text{focal}}; \phi_i^{\text{relation}}; \phi_i^{\text{peer}}; \phi_i^{\text{cross}}\right]. \tag{6}$$

3.4 SOURCE-TRAINED RISK SCORING

We map the representation to focal risk using L1-regularized logistic regression:

$$\hat{s}_i^{\text{SPAR}} = \sigma\left(\beta_0 + \boldsymbol{\beta}^\top \phi_i^{\text{SPAR}}\right). \tag{7}$$

During training, we assign equal total weight to each model–dataset domain. We then keep the feature construction and fitted coefficients fixed across all evaluation settings. The structure-free SPAR-SCALAR method retains focal uncertainty, mean focal–peer disagreement, and percentile-rank alignment, but removes peer–peer geometry, reliability structure, and their interactions. Exact feature definitions, fitting details, and the scalar score are provided in Sections B and D.

4 EXPERIMENTS

Our experiments ask four questions: whether SPAR improves selective-risk ordering across focal-model families and datasets; whether the source-trained scorer transfers under simultaneous model-family and dataset shift; whether answer relations explain the gain beyond learned edge-free controls; and whether the gains persist across correctness definitions and practical deployment constraints.

4.1 EXPERIMENTAL SETUP

Source development. We develop SPAR on VQAv2 and A-OKVQA (Goyal et al., 2017; Schwenk et al., 2022). Gemma-3-4B, Phi-4-Multimodal-5.6B, and InternVL3-8B rotate between focal and peer roles (Gemma Team, 2025; Abouelenin et al., 2025; Zhu et al., 2025). Representation selection, mechanism analysis, and ablations use only these source datasets.

Main benchmark. The main comparison covers MMBench (Liu et al., 2024), MMMU (Yue et al., 2024), and MMStar (Chen et al., 2024), rotating Gemma-3, Phi-4MM, InternVL3, GLM-4.1V (GLM-V Team, 2025), and Llama-3.2 (Grattafiori et al., 2024) through the focal role. Each focal VLM is paired with three peers under the fixed rotation rule in Section B; the source-trained scorer and peer-selection rule remain unchanged throughout.

Joint-shift transfer. Simultaneous model-family and dataset transfer applies the source-trained scorer to GLM-4.1V-9B-Thinking and Llama-3.2-11B-Vision-Instruct on the unseen TextVQA (Singh et al., 2019) and VizWiz (Gurari et al., 2018) datasets. Gemma-3, Phi-4MM, and InternVL3 serve as the peer models. Dataset construction, model checkpoints, and inference settings are detailed in Section B.

Baselines. We use length-normalized Sequence Likelihood (SL(norm)) (Malinin & Gales, 2021; Lin et al., 2024a) as the focal-only baseline. We additionally compare against Answer Entropy (Liu et al., 2026), Variation Ratios (Gal et al., 2017), KLE (Nikitin et al., 2024), SGD_s (Li et al., 2025), Majority Voting and SCoOP (Yu et al., 2026), and Consensus Entropy (Zhang et al., 2026b). SPAR-SCALAR is our structure-free ablation, retaining the same focal outputs and peer-call budget while compressing the peer relations into mean disagreement. Complete definitions and implementation details are provided in Section B.

All tunable baselines are selected exclusively on the source development data under the same macro-AURC objective and frozen before evaluation. Separately, SPAR w/o Relations uses the same source data, L1-regularized logistic learner, and model-selection protocol as SPAR, while excluding structured relation features beyond scalar focal–peer disagreement. It retains focal confidence, mean focal–peer disagreement, and peer-risk statistics, but removes peer–peer geometry, focal isolation, peer ambiguity, and answer-pattern indicators.

Metrics and comparison protocol. All methods rank the same fixed focal answers. Deterministic multi-VLM methods share the same four model calls as SPAR, while repeated-sampling methods are evaluated at multiple sampling budgets. Generation time quantifies the corresponding inference cost.

The primary metric is area under the risk–coverage curve (AURC), which summarizes selective risk over all coverage levels. Macro averages weight each family–dataset setting equally. Metric and correctness definitions are provided in Section B.6. Statistical uncertainty is estimated with paired bootstrap resampling.

Targeted robustness checks test two aspects of robustness. First, we vary the VQA correctness definition. Second, we test whether percentile-rank alignment depends on statistics from the full target set by replacing the full-target rank reference with source-only, causal-online, or fixed-buffer estimates. All checks use the same joint-shift focal outputs; complete configurations are reported in the appendix.

4.2 Main Results and Efficiency

Table 1 presents the broad fixed-focal comparison across five VLMs and three datasets. SPAR achieves the lowest average AURC on all three datasets and leads the broad majority of individual focal-model settings. When averaged equally over the displayed model–dataset settings, it reduces macro-AURC from 0.2305 for the strongest external competitor, Consensus Entropy, to 0.2211, corresponding to a 4.06% reduction. The gains extend to stronger sampling-based estimators: SPAR improves over SCoOP, KLE, and SGD_s by 18.32%, 39.99%, and 33.27%, respectively, while using only four model calls.

The measured efficiency comparison in Figure 1 demonstrates that this advantage does not rely on repeated sampling. On the shared evaluation set, SPAR delivers the largest macro-AURC reduction among the compared methods and a 10.1$\times$ generation-time speedup over SCoOP at its largest sampling budget (Table 5).

4.3 Transfer and Generalization

The same source-trained scorer also transfers when both the focal-model family and dataset change. As reported in Table 6, SPAR reduces macro-AURC from 0.2234 to 0.2076, a 7.05% improvement, with gains in every unseen model–dataset setting. As shown in Figure 3a, the advantage holds throughout the deployment range: relative to Consensus Entropy, SPAR reduces selective risk by 18.6%, 10.3%, and 3.2% at 25%, 50%, and 75% coverage, respectively. The same comparison raises error AUROC from 0.8610 to 0.8872.

Table 1: Setting-wise fixed-focal failure ranking across three datasets. Each model column specifies the focal VLM, whose three peers are selected as the first three remaining models in the fixed order of Gemma-3, Phi-4MM, InternVL3, Llama-3.2, and GLM-4.1V. **Bold** and underlined entries indicate the best and second-best results, respectively.

Method	Calls	Gemma-3	GLM-4.1V	InternVL3	Llama-3.2	Phi-4MM	Avg.
			MMBench				
SL(norm)	1	0.067	0.105	0.034	0.101	0.026	0.067
Answer Entropy	10	0.224	0.072	0.065	0.113	0.060	0.107
Variation Ratios	10	0.224	0.071	0.065	0.114	0.061	0.107
KLE	10	0.225	0.103	0.086	0.195	0.094	0.141
SGD_s	10	0.224	0.072	0.066	0.125	0.061	0.109
Majority Voting	40	0.065	0.056	0.052	0.108	0.060	0.068
SCoOP	40	0.045	0.043	0.038	0.089	0.042	0.051
Consensus Entropy	4	0.040	0.034	0.026	0.058	0.025	0.037
SPAR (*Scalar*)	4	0.042	0.038	0.026	0.059	0.025	0.038
SPAR (*Ours*)	4	**0.039**	**0.025**	**0.023**	**0.053**	**0.024**	**0.033**
			MMMU				
SL(norm)	1	0.424	0.462	0.293	0.525	0.272	0.395
Answer Entropy	10	0.570	0.474	0.326	0.591	0.310	0.454
Variation Ratios	10	0.570	0.478	0.326	0.594	0.314	0.456
KLE	10	0.578	0.555	0.431	0.639	0.435	0.528
SGD_s	10	0.570	0.504	0.326	0.613	0.325	0.468
Majority Voting	40	0.433	0.426	0.370	0.541	0.377	0.430
SCoOP	40	0.368	0.419	0.339	0.522	0.333	0.396
Consensus Entropy	4	0.363	0.350	0.297	0.436	0.287	0.347
SPAR (*Scalar*)	4	0.363	0.358	0.289	0.444	0.275	0.346
SPAR (*Ours*)	4	**0.359**	**0.345**	**0.275**	**0.430**	**0.265**	**0.335**
			MMStar				
SL(norm)	1	0.407	0.442	**0.175**	0.503	0.263	0.358
Answer Entropy	10	0.517	0.446	0.256	0.507	0.327	0.411
Variation Ratios	10	0.517	0.445	0.256	0.508	0.327	0.411
KLE	10	0.519	0.484	0.298	0.525	0.361	0.437
SGD_s	10	0.517	0.463	0.262	0.515	0.328	0.417
Majority Voting	40	0.407	0.461	0.316	0.445	0.335	0.393
SCoOP	40	0.350	0.461	0.286	0.425	0.299	0.364
Consensus Entropy	4	**0.331**	0.366	0.217	0.372	0.256	0.308
SPAR (*Scalar*)	4	0.337	0.373	0.199	0.384	0.254	0.309
SPAR (*Ours*)	4	0.334	**0.332**	0.214	**0.351**	**0.248**	**0.296**

Table 2 and Figure 3b extend this result across 4 model families and 7 focal VLMs spanning 4B to 90B, including the Qwen2.5-VL, Qwen3-VL, Gemma 3, and Llama 3.2 families (Bai et al., 2025a;b; Gemma Team, 2025; Grattafiori et al., 2024). The frozen SPAR scorer obtains the lowest model-level average AURC for every focal VLM and reduces aggregate AURC by 6.71% relative to the strongest external comparison. These results establish consistent scale generalization across heterogeneous VLM families from 4B to 90B parameters.

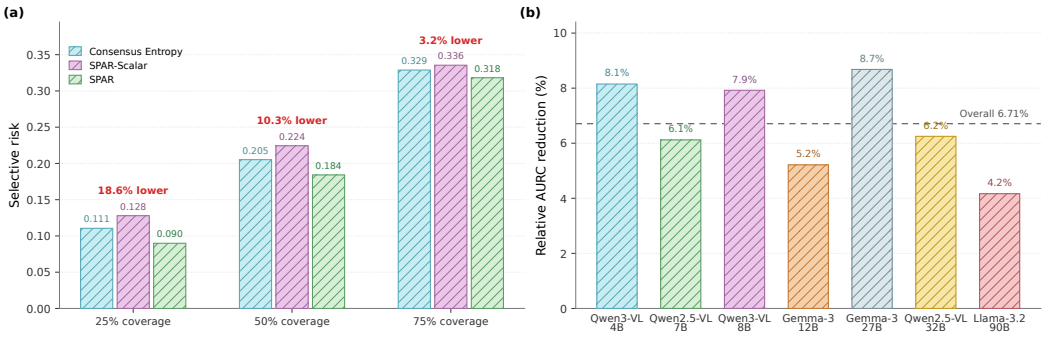


Figure 3: **Selective prediction and scale generalization.** **(a)** SPAR reduces selective risk at three joint-shift operating points. **(b)** SPAR improves AURC across seven 4B–90B focal VLMs; the dashed line denotes the aggregate reduction over Consensus Entropy.

Table 2: Scale transfer across seven focal VLMs from four model families, with model sizes ranging from 4B to 90B. Entries report AURC. **Bold** and underlined entries indicate the best and second-best results, respectively. SPAR achieves the best model-level average in every block.

		AURC ↓					
Focal VLM	Method	MMBench	MMMU	MMStar	TextVQA	VizWiz	Avg.
Qwen3-VL-4B	SL(norm)	0.0263	0.2868	0.2162	0.1555	**0.2379**	0.1845
	Consensus Entropy	0.0200	0.2740	0.2250	0.1032	0.2614	0.1767
	SPAR (*Ours*)	**0.0189**	**0.2610**	**0.1983**	**0.0950**	0.2382	**0.1623**
Qwen2.5-VL-7B	SL(norm)	0.0294	0.3318	0.2386	0.1612	0.2702	0.2062
	Consensus Entropy	0.0245	0.3020	0.2296	0.1062	0.2692	0.1863
	SPAR (*Ours*)	**0.0232**	**0.2865**	**0.2223**	**0.0974**	**0.2449**	**0.1749**
Qwen3-VL-8B	SL(norm)	**0.0154**	0.2504	0.1975	0.1580	**0.2211**	0.1685
	Consensus Entropy	0.0155	0.2553	0.2039	0.0985	0.2473	0.1641
	SPAR (*Ours*)	0.0155	**0.2397**	**0.1820**	**0.0914**	0.2271	**0.1511**
Gemma-3-12B	SL(norm)	0.0380	0.3029	0.3118	0.2088	0.4708	0.2665
	Consensus Entropy	0.0293	0.2818	0.2739	0.1387	0.3793	0.2206
	SPAR (*Ours*)	**0.0284**	**0.2669**	**0.2703**	**0.1329**	**0.3471**	**0.2091**
Gemma-3-27B	SL(norm)	0.0297	0.2695	0.2959	0.2650	0.5246	0.2769
	Consensus Entropy	0.0249	0.2651	0.2660	0.1406	0.3864	0.2166
	SPAR (*Ours*)	**0.0227**	**0.2436**	**0.2350**	**0.1321**	**0.3555**	**0.1978**
Qwen2.5-VL-32B	SL(norm)	0.0163	**0.2101**	0.1994	0.1481	**0.2339**	0.1616
	Consensus Entropy	0.0159	0.2271	0.2104	0.1059	0.2549	0.1628
	SPAR (*Ours*)	**0.0156**	0.2118	**0.1964**	**0.0984**	0.2352	**0.1515**
Llama-3.2-90B	SL(norm)	0.0402	0.3046	0.3018	0.1559	0.3684	0.2342
	Consensus Entropy	0.0293	0.2791	0.2766	0.1025	0.2844	0.1944
	SPAR (*Ours*)	**0.0265**	**0.2726**	**0.2581**	**0.0990**	**0.2752**	**0.1863**

We further evaluate SPAR on held-out DetailCaps (Dong et al., 2024), a detailed image-captioning benchmark, after fitting the risk scorer only on source captioning datasets. As shown in Table 7, SPAR achieves the best aggregate AURC, error AUROC, and rank correlation among all deterministic comparisons, reducing AURC from 0.4351 for Consensus Entropy to 0.4277 (1.70%), extending its effectiveness to long-form captions.

4.4 Contribution of Structured Relations

The transfer gain persists under matched learner-capacity controls. As shown in Table 3A, SPAR achieves 0.2076 AURC, improving over the matched w/o Relations control by 6.35% and substantially outperforming the learned scalar control. This comparison holds learner capacity nearly fixed: SPAR and this control both use 50 candidate coordinates and retain 33 and 31 active coefficients, respectively. By controlling for feature count and source supervision, this comparison isolates the improvement brought by our method.

The matched comparison also improves error detection: SPAR raises error AUROC from 0.8618 to 0.8872, a gain of 0.0254. Paired bootstrap analysis confirms that the AURC improvement is consistently positive. Together, these results show that relation-aware evidence adds predictive signal beyond scalar consensus and a w/o Relations control with the same candidate dimension and a closely matched active-coefficient count.

4.5 Ablation and Robustness

The leave-one-group-out results in Table 10 quantify the four evidence groups used by SPAR. On the development evaluation, removing any group increases macro-AURC. Removing the peer-reliability group causes the largest increase (2.52%), followed by answer-relation geometry (2.09%), focal

Table 3: **Relational structure and practical rank alignment under joint model-family and dataset shift.** Panel (a) shows that SPAR outperforms its w/o Relations control, which has the same candidate dimension and a closely matched active-coefficient count. Panel (b) shows that the advantage persists with full-target, source-only, causal-online, and fixed-buffer rank references. In Panel (a), bold and underlining mark the best and second-best AURC; in Panel (b), bold marks the better method in each row.

(a) Relational-structure control

Variant	Candidate coords.	Active coeff.	AURC ↓
SPAR (*Scalar*)	2	2	0.2287
SPAR w/o Relations	50	31	<u>0.2217</u>
SPAR (*Ours*)	50	33	**0.2076**

(b) Rank-reference control

Rank reference	Consensus Entropy	SPAR	Reduction (%) ↑
Full target	0.2234	**0.2076**	7.05
Source-only (model-specific)	0.2199	**0.2049**	6.84
Causal online (128)	0.2212	**0.2058**	6.99
Fixed buffer (128)	0.2235	**0.2075**	7.18

confidence (1.76%), and cross-evidence interactions (1.43%). These results support the contribution of each group within the source-selected representation.

As shown in Table 3B, SPAR retains 6.84%–7.18% AURC reductions with source-only, causal-online, and fixed-buffer rank references, demonstrating robust deployment without full-target statistics.

The advantage persists across four VQA correctness rules—full consensus, consensus score ≥ 0.6, any nonzero consensus credit, and continuous soft risk—with 7.05%–10.42% AURC reductions over Consensus Entropy (Table 4).

Table 4: **Robustness to correctness definitions.** SPAR achieves the lowest AURC under all four evaluation rules. Bold and underlined entries indicate the best and second-best results, respectively.

Correctness definition	SL(norm)	Consensus Entropy	SPAR	Reduction vs. CE (%) ↑
Full consensus	0.3366	<u>0.2234</u>	**0.2076**	7.05
Consensus ≥ 0.6	0.2534	<u>0.1508</u>	**0.1374**	8.93
Any consensus credit	0.2064	<u>0.1196</u>	**0.1071**	10.42
Continuous soft risk	0.2584	<u>0.1574</u>	**0.1437**	8.72

Further controls show that SPAR improves with one, two, and three peers, retains 4.50%–5.73% mean relative AURC gains under 75% joint answer–confidence corruption, and exhibits zero representation drift under all 16 monotone rescalings (Tables 11 and 12). Together, these results demonstrate that the advantage from structured answer relations persists across correctness rules, rank-reference choices, peer budgets, and substantial evidence corruption.

5 CONCLUSION

We introduced SPAR, a confidence-aware relational framework for selective prediction with heterogeneous VLMs. By combining the decomposition of focal disagreement into focal isolation and peer ambiguity with relative peer confidence and percentile-rank alignment, SPAR preserves answer structure that scalar consensus discards and converts it into a transferable risk ordering. Extensive experiments show consistent improvements over peer-based and repeated-sampling alternatives, including under simultaneous model-family and dataset shift and across focal models spanning 4B to 90B parameters. Capacity-matched controls isolate the contribution of relational evidence, while correctness and rank-reference analyses show that the advantage persists under practical evaluation and deployment choices. Together, these results establish structured answer relations as an effective and transferable reliability signal for fixed-focal selective prediction.

AI USE STATEMENT

Generative AI tools were used only for limited language polishing and LaTeX formatting assistance during manuscript preparation. The research content, experimental results, and figures were produced by the authors. The authors reviewed all suggested edits and take full responsibility for the final manuscript.

REFERENCES

Abdelrahman Abouelenin et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. arXiv preprint arXiv:2503.01743.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019. doi: 10.1109/ICCV.2019.00904.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report, 2025a. arXiv preprint arXiv:2502.13923.

Shuai Bai et al. Qwen3-VL technical report, 2025b. arXiv preprint arXiv:2511.21631.

Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. Post-hoc probabilistic vision-language models. In *International Conference on Learning Representations*, 2026.

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 *Advances in Neural Information Processing Systems*, volume 37, 2024.

Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24049–24059, 2023.

Hao Dong, Moru Liu, Jian Liang, Eleni Chatzi, and Olga Fink. To trust or not to trust your vision-language model's prediction, 2025. arXiv preprint arXiv:2505.23745.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption, 2024. arXiv preprint arXiv:2405.19092.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192, 2017.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. ImageInWords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 93–127, 2024. doi: 10.18653/v1/2024.emnlp-main.6.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Gemma Team. Gemma 3 technical report, 2025. arXiv preprint arXiv:2503.19786.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. arXiv preprint arXiv:2507.01006.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. arXiv preprint arXiv:2407.21783.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

Nhat Hoang, Minh Vu, My T. Thai, and Manish Bhattarai. PAS: Prelim attention score for detecting object hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18273–18283, 2026.

Tao Huang, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, and Liping Jing. Detecting misbehaviors of large vision-language models by evidential uncertainty quantification. In *International Conference on Learning Representations*, 2026.

Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10854–10863, 2024.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 6067–6085, 2026. doi: 10.18653/v1/2026.eacl-long.287.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Marc Lafon, Yannis Karmim, Julio Silva-Rodríguez, Paul Couairon, Clément Rambour, Raphael Fournier-Sniehotta, Ismail Ben Ayed, Jose Dolz, and Nicolas Thome. ViLU: Learning vision-language uncertainties for failure prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17807–17817, 2025. doi: 10.1109/ICCV51701.2025.01654.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2789.

Zhaoye Li, Siyuan Shen, Wenjing Yang, Ruochun Jin, Huan Chen, Ligong Cao, and Jing Ren. Enhancing uncertainty quantification in large language models through semantic graph density. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 2537–2551, 2025.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10351–10368, 2024a. doi: 10.18653/v1/2024.emnlp-main.578.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024b.

Litian Liu, Reza Pourreza, Sunny Panchal, Apratim Bhattacharyya, Yubing Jian, Yao Qin, and Roland Memisevic. Enhancing hallucination detection through noise injection. In *International Conference on Learning Representations*, 2026.

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 *Proceedings of the European Conference on Computer Vision*, pp. 216–233, 2024. doi: 10.1007/978-3-031-72658-3_13.

Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Yang Nan, Pengfei He, Ravi Tandon, and Han Xu. Can multiple responses from an LLM reveal the sources of its uncertainty? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15551–15569, 2025. doi: 10.18653/v1/2025.findings-emnlp.841.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0283.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264.

Aditya Sarkar, Yi Li, Jiacheng Cheng, Shlok Mishra, and Nuno Vasconcelos. Leveraging data to say no: Memory augmented plug-and-play selective prediction. In *International Conference on Learning Representations*, 2026.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision*, pp. 146–162, 2022. doi: 10.1007/978-3-031-20074-8_9.

Jie Shi, Xiaodong Yue, Wei Liu, Yufei Chen, and Feifan Dong. Not all inconsistency is equal: Decomposing LVLM uncertainty into belief divergence and belief conflict. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25339–25347, 2026. doi: 10.1609/aaai.v40i30.39727.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A dataset for image captioning with reading comprehension. In *Computer Vision – ECCV 2020*, pp. 742–758, 2020. doi: 10.1007/978-3-030-58536-5_44.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Guangzhi Sun, Potsawee Manakul, Adian Liusie, Kunat Pipatanakul, Chao Zhang, Phil Woodland, and Mark Gales. CrossCheckGPT: Universal hallucination ranking for multimodal foundation models. In *NeurIPS 2024 Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

Selim Furkan Tekin, Yichang Xu, Gaowen Liu, Ramana Rao Kompella, Margaret L. Loper, and Ling Liu. Vision verification enhanced fusion of VLMs for efficient visual reasoning, 2026. arXiv preprint arXiv:2603.12669.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, and Maxim Panov. UNCERTAINTY-LINE: Length-invariant estimation of uncertainty for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7881–7908, 2025. doi: 10.18653/v1/2025.emnlp-main.400.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *Proceedings of the European Conference on Computer Vision*, pp. 148–166, 2022. doi: 10.1007/978-3-031-20059-5_9.

Tobias Jan Wieczorek, Nathalie Daun, Mohammad Emtiyaz Khan, and Marcus Rohrbach. Variational visual question answering for uncertainty-aware selective prediction. *Transactions on Machine Learning Research*, 2026.

Chung-En Johnny Yu, Brian Jalaian, and Nathaniel D. Bastian. SCoOP: Semantic consistent opinion pooling for uncertainty quantification in multiple vision-language model systems. In *ICLR 2026 Workshop on Agentic AI in the Wild: From Hallucinations to Reliable Autonomy*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, 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 Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Jusheng Zhang, Yijia Fan, Kaitong Cai, Jing Yang, Jiawei Yao, Jian Wang, Guanlong Qu, Ziliang Chen, and Keze Wang. Why keep your doubts to yourself? trading visual uncertainties among vision-language models. In *International Conference on Learning Representations*, 2026a.

Yulong Zhang, Tianyi Liang, Erfei Cui, Guoqing Wang, Xu Guo, Chenhui Li, and Gongshen Liu. Consensus entropy: Harnessing multi-vlm agreement for self-verifying and self-improving OCR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11643–11653, 2026b.

Xingtao Zhao, Hao Peng, Dingli Su, Xianghua Zeng, Chunyang Liu, Jinzhi Liao, and Philip S. Yu. SeSE: Black-box uncertainty quantification for large language models based on structural information theory, 2025. arXiv preprint arXiv:2511.16275; accepted at UAI 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. arXiv preprint arXiv:2504.10479.

A Appendix Overview

The appendix provides reproducibility details and additional analyses. Section B describes datasets, models, parsing, baselines, statistics, and computational records. Section C reports detailed transfer, selective, robustness, and format-failure analyses. Section D gives the complete feature specification. Section E discusses limitations and future directions.

B Reproducibility and Experimental Details

B.1 Datasets and Sample Construction

Source development uses VQAv2 and A-OKVQA. The held-out joint family–dataset transfer uses 5,000 TextVQA validation examples and all 4,319 available VizWiz validation examples per focal model. Family–dataset settings are kept separate during evaluation, and macro results assign equal weight to each setting rather than to each example. Repeated questions associated with one image remain in the same bootstrap cluster.

B.2 Models, Prompts, and Decoding

Development rotates Gemma-3, Phi-4MM, and InternVL3 through the focal role. The primary transfer fixes these three models as peers and evaluates GLM-4.1V and Llama-3.2 as unseen focal families. For the broad comparison, the ordered model list is Gemma-3, Phi-4MM, InternVL3, Llama-3.2, and GLM-4.1V; for each focal VLM, the first three distinct models other than the focal serve as peers. Unless otherwise specified for repeated-sampling baselines, SPAR and deterministic peer-based comparisons use one deterministic generation per model and example. Each model uses a fixed prompt and generation configuration throughout evaluation.

B.3 Answer Parsing and Correctness

Answers are normalized with the standard VQA text normalization before correctness and token-level F1 similarities are computed. A focal answer is counted as correct only when its VQA consensus score equals one. Partial-credit, unparsed, and generation-limit outputs are treated as failures in

the primary metric. Parser identity and generation-limit status are retained as separate fields so that format failures can be analyzed without reconstructing an answer from incomplete reasoning.

B.4 Score Construction and Scorer Fitting

The generated-answer risk u_i^k is defined from length-normalized likelihood in Equation (3). Normalized Sequence Likelihood is the corresponding focal-only ordering.

For confidence-aware disagreement, increasing the weight of peer k changes Equation (5) according to

$$\frac{\partial D_{w,i}}{\partial c_i^k} = \frac{d_i^k - D_{w,i}}{\sum_{\ell=1}^{K} c_i^\ell}, \tag{8}$$

so confident opposition and confident support affect the representation in opposite directions.

The SPAR risk scorer is L1 logistic regression. Nested, grouped source validation selects the regularization strength from a prespecified grid; the selected final scorer uses $C = 1$, the `liblinear` solver, at most 1,000 iterations, and random seed 20260727. Each source model–dataset domain receives equal total weight. The 50-dimensional feature order and all fitted coefficients are fixed before target labels are accessed.

B.5 Baselines and Hyperparameters

Fusion weights for tunable external comparisons are selected by source-only validation over a prespecified candidate grid. All compared scores rank the same immutable focal outputs. SPAR (*Scalar fusion*) retains focal risk and mean focal–peer disagreement:

$$\hat{s}_i^{\text{scalar}} = R_{G_0}(u_i^0) + \alpha R_G(D_i). \tag{9}$$

The learned scalar control in Table 3A uses the same two coordinates, $R_{G_0}(u_i^0)$ and $R_G(D_i)$, but learns their coefficients with the same source-only sparse-logistic protocol as SPAR. Consensus Entropy replaces the disagreement term with the focal answer's mean normalized character-edit distance to the peer panel and uses the same target-rank interface. Restricted learned controls use the same source protocol as SPAR but retain only their specified feature groups.

The deterministic peer-based comparison uses one focal output and the same three peer outputs per example. No method in this comparison uses repeated sampling, an auxiliary judge, answer replacement, or routing. SL(norm) uses only the focal output and therefore has a lower call budget. Systems that alter the evaluated answer or require additional generations must be reported separately from this fixed-focal comparison.

The broad comparison in Table 1 evaluates MMBench, MMMU, and MMStar across five focal VLMs. It includes SL(norm); Answer Entropy, Variation Ratios, KLE, and SGD_s at $k = 10$; Majority Voting and SCoOP at $k = 10$; normalized-edit Consensus Entropy; SPAR-SCALAR; and SPAR. SCoOP preserves normalized per-model entropy, inverse-entropy weighting, linear opinion pooling, and fused entropy; its MCQ option space is replaced with exact clusters after standard VQA answer normalization. Majority Voting uses the same clusters and preserves the model-level vote and local-probability tie break.

The efficiency comparison in Figure 1 and Table 5 uses the same immutable focal answers on 1,000 examples from each of TextVQA and VizWiz for each focal family, selected by a prespecified hash order before target-label access. It visualizes nine configurations: SL(norm), Answer Entropy at $k = 10$, Majority Voting at $k = 10$, normalized-edit Consensus Entropy, SCoOP at $k \in \{3, 5, 10\}$, SPAR-SCALAR, and SPAR. Answer Entropy uses 10 focal generations; SCoOP uses 12, 20, and 40 total VLM calls at the three sampling budgets, and Majority Voting uses 40. Consensus Entropy, SPAR-SCALAR, and SPAR each use one focal and three peer generations, while SL(norm) uses only the focal generation. Every configuration ranks the same immutable focal answers and is evaluated without answer replacement. This comparison is reported separately from the primary evaluation on 5,000 TextVQA and 4,319 VizWiz examples.

B.6 Metrics and Statistical Analysis

For n examples ordered by increasing predicted risk, let $\mathrm{risk}(m)$ be the error rate among the first m accepted examples. The primary metric is

$$\mathrm{AURC} = \frac{1}{n} \sum_{m=1}^{n} \mathrm{risk}(m), \tag{10}$$

where lower is better. Secondary metrics are error AUROC and risk at 25%, 50%, and 75% coverage. Score ties are resolved by immutable sample ID. Paired uncertainty estimates use 10,000 resamples of image clusters within each fixed family–dataset setting, followed by an equal-weight average across settings.

All representation choices, hyperparameters, baselines, and analysis procedures were finalized using source data. Target correctness labels were used only after risk scores had been produced.

B.7 Computational Cost

All efficiency measurements were collected on AMD Instinct MI210 GPUs. Table 5 reports generation time, GPU-hours, model-size-weighted calls, and peak memory under the shared evaluation protocol. Model-size-weighted calls account for the different checkpoint sizes used by each method, while runtime and memory are measured directly.

Table 5: Compute accounting for the shared efficiency comparison.

Method	Calls	Model-weighted calls (B)	GPU s/ example	GPU-hours	Peak VRAM (GiB)
SL(norm)	1	10.48	5.48	6.09	26.80
Answer Entropy	10	104.81	55.73	61.92	26.80
Majority Voting	40	286.70	100.31	111.46	26.80
SCoOP ($k = 3$)	12	86.01	30.64	34.05	26.80
SCoOP ($k = 5$)	20	143.35	50.14	55.71	26.80
SCoOP ($k = 10$)	40	286.70	100.31	111.46	26.80
Consensus Entropy	4	28.67	9.96	11.07	26.80
SPAR (*Scalar*)	4	28.67	9.96	11.07	26.80
SPAR (*Ours*)	4	28.67	9.96	11.07	26.80

SPAR uses one focal generation and three peer generations per example. Graph construction and risk scoring operate on the resulting answers, leaving VLM generation as the dominant computational cost.

C Additional Experimental Results

C.1 Primary Joint-Transfer Results

Table 6: AURC under joint model-family and dataset transfer. **Bold** and underlined entries indicate the best and second-best results, respectively.

	GLM-4.1V		Llama-3.2		
Method	TextVQA	VizWiz	TextVQA	VizWiz	Macro
SL(norm)	0.1917	0.5326	0.1618	0.4602	0.3366
Consensus Entropy	<u>0.1043</u>	<u>0.3544</u>	<u>0.1119</u>	0.3229	<u>0.2234</u>
SPAR (*Scalar fusion*)	0.1115	0.3887	0.1147	<u>0.3219</u>	0.2342
SPAR (*Ours*)	**0.0940**	**0.3194**	**0.1070**	**0.3101**	**0.2076**

Across both unseen focal families and both target datasets, SPAR improves over Consensus Entropy, with relative AURC reductions ranging from 3.95% to 9.88%.

C.2 LONG-FORM CAPTIONING TRANSFER

We additionally test whether the relation-aware risk architecture extends beyond short-form VQA. The same risk scorer is fitted on TextCaps, NoCaps, and IIW-400 source captioning data (Sidorov et al., 2020; Agrawal et al., 2019; Garg et al., 2024), then frozen before evaluation on a held-out split of DetailCaps (Dong et al., 2024). Gemma-3-4B, Phi-4-Multimodal-5.6B, InternVL3-8B, and Idefics2-8B rotate through the focal role, with the other three models serving as peers (Gemma Team, 2025; Abouelenin et al., 2025; Zhu et al., 2025; Laurençon et al., 2024). Each method ranks the same detailed focal captions. Within each focal-model setting, caption quality is the mean of empirical CIDEr and ROUGE-L percentiles, and continuous risk is one minus this quality. AURC averages continuous risk over accepted prefixes, error AUROC treats the lowest-quality quartile as failure, and Spearman's ρ correlates predicted risk with continuous risk. UNCERTAINTY-LINE corrects focal likelihood for output-length effects (Vashurin et al., 2025). The semantic estimators use one deterministic answer from each VLM, matching SPAR's four-call interface.

Table 7: **Long-form captioning transfer on DetailCaps.** Results are averaged across four focal VLMs. Bold and underlined entries indicate the best and second-best results, respectively.

Method	AURC ↓	Error AUROC ↑	Spearman ρ ↑
SL(norm)	0.465	0.582	0.173
UNCERTAINTY-LINE	0.471	0.577	0.151
Semantic Entropy	0.503	0.493	-0.018
KLE	0.499	0.508	0.007
SGD_s	0.499	0.507	0.008
Consensus Entropy	<u>0.435</u>	<u>0.644</u>	<u>0.285</u>
SPAR (*Ours*)	**0.428**	**0.672**	**0.333**

As shown in Table 7, SPAR achieves the best aggregate result on all three metrics. It reduces AURC by 1.70% relative to Consensus Entropy, with a paired 95% interval of [0.97%, 2.42%], and improves upon that strongest deterministic comparison for 3 of the 4 focal VLMs. This complementary result demonstrates that SPAR's relational evidence remains effective after source-caption fitting and transfer to an unseen long-form captioning domain.

C.3 STATISTICAL SIGNIFICANCE AND SOURCE-RESAMPLING STABILITY

We examine uncertainty from finite evaluation samples and variation induced by source-data resampling.

Paired statistical significance. Holding the fixed source-trained scorer constant, we resample evaluation image clusters within each model–dataset setting. As shown in Table 8, the paired 95% intervals remain above zero on joint family–dataset transfer, the broad MCQ comparison, and scale transfer, establishing significant AURC reductions across all three evaluation regimes.

Source-resampling stability. We separately refit the risk scorer using five independently seeded, dataset-stratified source image-cluster resamples while keeping the VLM predictions, percentile-rank alignment, and evaluation data fixed. The 72B–90B evaluation uses a fixed peer panel of Gemma-3-4B, Phi-4-Multimodal-5.6B, and InternVL3-8B. The refit columns of Table 8 summarize the resulting variation. Every refit improves over Consensus Entropy across the reported evaluation scopes, supporting stability under source-data resampling.

C.4 CONTROLLED ANALYSES UNDER JOINT TRANSFER

Learned capacity controls. Table 9 compares SPAR with learned scalar, dimension-matched, nonlinear edge-free, and w/o Relations controls. The w/o Relations model uses the same source-training and model-selection protocol and 50-coordinate candidate dimension, selecting 31 active coefficients, compared with 33 for SPAR. Joint-transfer AURC falls from 0.2217 to 0.2076, a relative reduction of 6.35%, while error AUROC rises from 0.8618 to 0.8872. The paired AURC gain has a 95% interval of [0.0127, 0.0154], showing that the transfer advantage is not explained by feature count, coefficient count, or source supervision alone.

Table 8: Statistical significance and source-resampling stability of SPAR relative to Consensus Entropy. Fixed-scorer columns report relative AURC reductions and paired bootstrap confidence intervals. Refit columns summarize five source-data resamples as mean ± standard deviation. Dashes indicate unreported fixed-scorer statistics. The 72B–90B scope covers Qwen2.5-VL-72B-Instruct, InternVL3-78B, and Llama-3.2-90B-Vision-Instruct on MMBench, MMMU, MMStar, TextVQA, and VizWiz (15 model–dataset settings). The ScienceQA-IMG and POPE scope covers Gemma-3-4B, GLM-4.1V-9B-Thinking, InternVL3-8B, Llama-3.2-11B-Vision-Instruct, and Phi-4-Multimodal-5.6B (10 settings).

	Fixed scorer		Source-data refits	
Evaluation scope	Reduction (%) ↑	Paired 95% CI (%)	AURC ↓	Reduction (%) ↑
Joint family–dataset transfer	7.05	[6.19, 7.89]	0.2094 ± 0.0014	6.27 ± 0.64
Broad MCQ comparison	4.06	[3.12, 4.96]	0.2214 ± 0.0013	3.93 ± 0.56
Scale transfer (35 settings)	6.71	[6.33, 7.24]	0.1616 ± 0.0010	6.16 ± 0.57
72B–90B focal VLMs	–	–	0.1527 ± 0.0011	4.96 ± 0.68
ScienceQA-IMG and POPE transfer	–	–	0.0562 ± 0.0005	2.63 ± 0.81

Table 9: Learned controls for feature dimension, active-coefficient count, and learner capacity. Bold and underlining mark the best and second-best AURC.

Variant	Candidate coords.	Active coeff.	Joint-transfer AURC ↓
Learned scalar	2	2	0.2287
Edge-free linear	17	9	<u>0.2148</u>
Edge-free (dimension-matched)	50	13	0.2219
SPAR w/o Relations	50	31	0.2217
Edge-free nonlinear	17	–	0.2162
SPAR (*Ours*)	50	33	**0.2076**

Target rank references. Figure 4 compares nine full-target, source-only, finite-buffer, and online rank references across the joint-transfer settings. All 36 reductions over Consensus Entropy remain positive, including the ranges induced by buffer sampling and input order. Together with the representative AURC values in Table 3B, this result shows that SPAR retains its advantage without access to the complete target-unlabeled distribution.

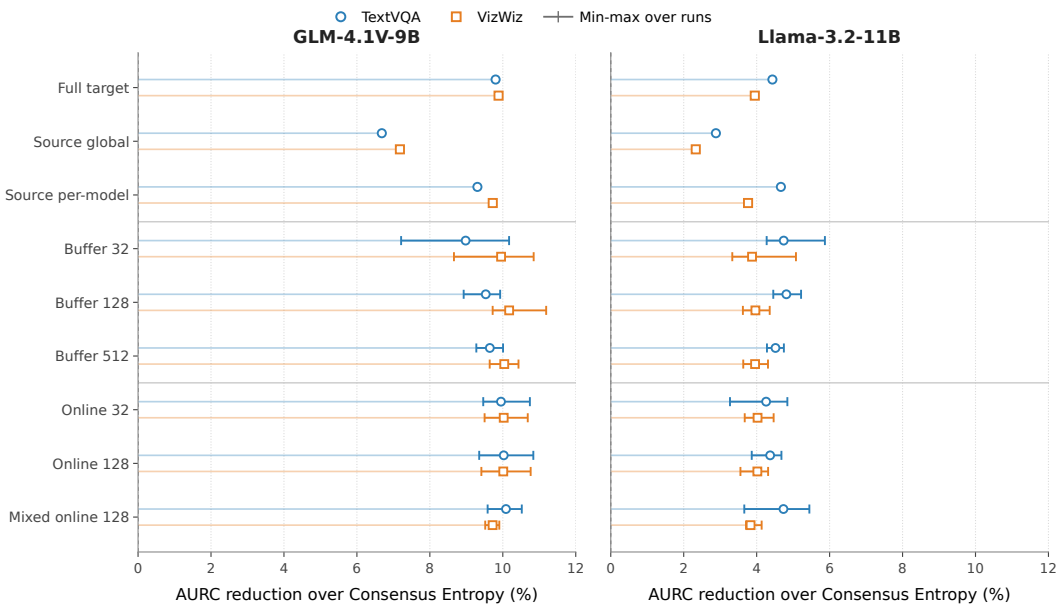


Figure 4: **Robustness to rank-reference estimation under joint transfer.** Relative AURC reductions of SPAR over Consensus Entropy across nine rank-reference schemes. Markers show deterministic estimates or means; error bars span five sampled buffers or 20 input orders. Positive reductions across both focal VLMs and target datasets demonstrate robustness to source-only, finite-buffer, and online rank references.

C.5 FEATURE-ABLATION DETAILS

Table 10 reports the leave-one-group-out component ablation summarized in Section 4.5.

Table 10: Component ablation of SPAR under leave-one-family-out development.

Variant	Macro-AURC↓	Relative increase
SPAR (*Ours*)	**0.2104**	—
w/o Cross-Evidence Interactions	0.2134	+1.43%
w/o Answer-Relation Geometry	0.2148	+2.09%
w/o Focal Confidence	0.2141	+1.76%
w/o Peer-Reliability Proxies	0.2157	+2.52%

C.6 CALL-BUDGET DETAILS

The standard representation uses three peer slots. In the reduced-budget evaluation, these slots are populated cyclically from one or two distinct peers, preserving the scorer input schema while varying the number of VLM calls. Table 11 reports the resulting comparison.

Table 11: Call-budget ablation over held-out families and peer panels. Relative AURC gain is measured against the stronger of SPAR (*Scalar*) and Consensus Entropy; brackets summarize the range across evaluations.

Distinct peers	Total calls	Evaluations	Mean gain	Improved
1	2	12	3.49% [-3.63, 8.40]	10
2	3	12	5.20% [-0.07, 9.39]	11
3	4	4	3.20% [0.86, 4.54]	4

C.7 PEER AND SCALE STRESS TESTS

Table 12 reports peer duplication, 75% joint corruption, and monotone-rescaling analyses.

Table 12: Robustness to peer perturbations and monotone rescaling. Joint corruption simultaneously perturbs answers and confidence on 75% of examples; brackets span four held-out families and three seeds. Strictly increasing transforms leave empirical-rank features unchanged.

Perturbation	Reported quantity	Result
Duplicate highest-confidence peer	Relative AURC gain	+4.29%
Duplicate lowest-confidence peer	Relative AURC gain	+5.22%
Corrupt one peer on 75% of examples	Mean relative AURC gain	+4.50% [+2.15, +8.23]
Corrupt two peers on 75% of examples	Mean relative AURC gain	+5.73% [+1.60, +10.90]
Strictly increasing scale transforms	Maximum rank-feature drift	0.0 (16 scenarios)

C.8 FORMAT-FAILURE SENSITIVITY

Table 13 tests whether parser or generation-limit failures explain the primary advantage. The reduction over Consensus Entropy persists within both focal families after excluding the relevant failure strata under equal-dataset macro aggregation.

Table 13: Format-failure sensitivity under joint model-family and dataset transfer. AURC is averaged equally across TextVQA and VizWiz.

Focal	Retained stratum	n	SPAR	CE	Reduction
GLM-4.1V	All outputs	9,319	0.2067	0.2293	9.87%
GLM-4.1V	Successfully parsed	8,979	0.1972	0.2211	10.78%
GLM-4.1V	No generation-limit hit	9,027	0.1988	0.2224	10.60%
Llama-3.2	All outputs	9,319	0.2086	0.2174	4.07%
Llama-3.2	No generation-limit hit	9,276	0.2071	0.2160	4.12%

D FEATURE SPECIFICATION

The SPAR representation is constructed from pairwise answer relations and likelihood-derived model-risk coordinates. Its four evidence groups correspond to the construction in Equation (6). This section specifies their fixed coordinate order. For compactness, $\mathrm{stat}_{\mathrm{mean,min,max,sd}}(\cdot)$ returns the listed population statistics in that order; analogous notation is used below for other ordered statistics.

Primitive answer relations. For normalized answers a and b, let $n_a(w)$ and $n_b(w)$ denote the multiplicities of token w, and let $N_a = \sum_w n_a(w)$ and $N_b = \sum_w n_b(w)$. Their token-count overlap is

$$o(a,b) = \sum_w \min\{n_a(w), n_b(w)\}, \qquad s(a,b) = \begin{cases} \dfrac{2o(a,b)}{N_a + N_b}, & N_a + N_b > 0, \\ 0, & N_a + N_b = 0. \end{cases} \tag{11}$$

This is equivalent to the harmonic mean of token precision $o(a,b)/N_a$ and recall $o(a,b)/N_b$ whenever the corresponding denominators are nonzero.

Write $s_k = s(a_i^0, a_i^k)$, $p_{k\ell} = s(a_i^k, a_i^\ell)$, and $d_k = 1 - s_k$. The mean focal disagreement D_i, focal isolation I_i, and peer ambiguity A_i are defined in Equation (2). Let $\tilde{a}_i^k$ denote the VQA-normalized answer and let n_b be the number of peers whose normalized answer is b. If $n_f = n_{\tilde{a}_i^0}$ and $n_1 \geq n_2$ are the two largest counts among answers other than $\tilde{a}_i^0$, with missing counts set to zero, then focal support, opposition support, and the opposition margin are

$$F_i = \frac{n_f}{K}, \qquad O_i = \frac{n_1}{K}, \qquad \Delta_i = \frac{n_1 - \max(n_f, n_2)}{K}. \tag{12}$$

For peer centrality, we use

$$m_k = \frac{1}{K-1}\sum_{\ell \neq k} p_{k\ell}, \qquad w_k = \frac{\exp\{3(m_k - \max_j m_j)\}}{\sum_j \exp\{3(m_j - \max_\ell m_\ell)\}}. \tag{13}$$

The medoid set is $\mathcal{M} = \arg\max_k m_k$. With peer-answer frequencies $\pi_b = n_b/K$, normalized vote entropy and centrality-weight entropy are

$$H_i^{\text{vote}} = -\frac{\sum_b \pi_b \log \pi_b}{\log K}, \qquad H_i^{\text{cent}} = -\frac{\sum_k w_k \log w_k}{\log K}. \tag{14}$$

Finally, the focal answer's mean normalized character-edit distance to the peers is

$$E_i = \frac{1}{K}\sum_{k=1}^{K} \frac{\text{Lev}(\tilde{a}_i^0, \tilde{a}_i^k)}{\max(|\tilde{a}_i^0|, |\tilde{a}_i^k|)}, \tag{15}$$

where the ratio is set to zero when both normalized strings are empty.

Focal confidence (1). The first group contains only the rank-aligned absolute focal risk,

$$\phi_i^{\text{focal}} = [R_{G_0}(u_i^0)]. \tag{16}$$

Answer-relation geometry (27). The 23 continuous coordinates are, in order,

$$\begin{aligned}\psi_i^{\text{relation}} = \big[&D_i, I_i, A_i, O_i, \Delta_i, F_i;\\ &\text{stat}_{\text{min,max,sd}}(\{s_k\}); \text{stat}_{\text{min,max,sd}}(\{p_{k\ell} : k < \ell\});\\ &H_i^{\text{vote}}; E_i; \text{stat}_{\text{min,max,median}}(\{d_k\});\\ &\text{stat}_{\text{min,max,sd}}(\{m_k\}); \textstyle\sum_k w_k d_k; |\mathcal{M}|^{-1}\sum_{k\in\mathcal{M}} d_k; H_i^{\text{cent}}\big].\end{aligned} \tag{17}$$

Four binary coordinates encode, respectively,

$$[\nVdash(F_i = 1), \nVdash(O_i = 1), \nVdash(0 < F_i < 1), \nVdash(F_i = 0 \wedge O_i < 1)], \tag{18}$$

corresponding to unanimous support, unanimous opposition, partial support, and fragmented opposition. Exactly one is active. Applying R_G coordinate-wise to ψ_i^{relation} and appending these indicators gives ϕ_i^{relation}.

Peer-reliability proxies (13). Let $r_k = R_{G_k}(u_i^k)$, $c_k = 1 - r_k$, and $\omega_k = c_k / \sum_j c_j$; if the denominator is zero, set $\omega_k = 1/K$. Let $\mathcal{S} = \{k : \tilde{a}_i^k = \tilde{a}_i^0\}$ and $\mathcal{O} = \{1, \ldots, K\} \setminus \mathcal{S}$, and define maxima over an empty set as zero. We also write $\mathcal{C}_{\max} = \arg\max_k c_k$ and $\mathcal{C}_{\min} = \arg\min_k c_k$. The ordered reliability vector is

$$\begin{aligned}\psi_i^{\text{peer}} = \big[&\text{stat}_{\text{mean,min,max,sd}}(\{r_k\}); \textstyle\sum_k \omega_k d_k; |\mathcal{C}_{\max}|^{-1}\sum_{k\in\mathcal{C}_{\max}} d_k; |\mathcal{C}_{\min}|^{-1}\sum_{k\in\mathcal{C}_{\min}} d_k;\\ &\textstyle\sum_{k\in\mathcal{O}} \omega_k; \max_{k\in\mathcal{S}} c_k; \max_{k\in\mathcal{O}} c_k; \max_{k\in\mathcal{O}} c_k - \max_{k\in\mathcal{S}} c_k;\\ &K^{-1}\textstyle\sum_k (c_k - \bar{c})(d_k - \bar{d}); \max_k(c_k d_k)\big].\end{aligned} \tag{19}$$

Thus this group contains both the peer-risk distribution and relation statistics conditioned on likelihood-derived peer-confidence proxies. Coordinate-wise alignment of ψ_i^{peer} under G gives ϕ_i^{peer}.

Cross-evidence interactions (9). Let hats denote aligned coordinates, e.g., $\widehat{D}_i = R_G(D_i)$, $\widehat{u}_i = R_{G_0}(u_i^0)$, $\widehat{D}_{w,i} = R_G(D_{w,i})$, and let $\widehat{O}_{w,i} = R_G(\sum_{k\in\mathcal{O}} \omega_k)$. Peer consistency is $C_i = 1 - R_G(A_i)$. The cross-evidence group is

$$\begin{aligned}\phi_i^{\text{cross}} = \big[&\widehat{D}_i C_i, \widehat{I}_i C_i, \widehat{O}_i C_i, \widehat{u}_i \widehat{I}_i, \widehat{u}_i \widehat{O}_i, \widehat{u}_i \widehat{A}_i;\\ &\widehat{D}_{w,i} C_i, \widehat{D}_{w,i}\widehat{u}_i, \widehat{O}_{w,i}\widehat{u}_i\big].\end{aligned} \tag{20}$$

These products test whether relational evidence becomes more informative when the focal model is uncertain, the peers are internally coherent, or the opposing peers receive high likelihood-derived confidence.

All 37 continuous base coordinates use the average-tie percentile in Equation (4); answer-pattern indicators remain binary and interactions are formed only after alignment. The intercept is not counted as a feature. The estimator is L1 logistic regression with $C = 1$; its 50 coefficients contain 33 nonzero entries. The feature order and fitted coefficients are fixed before target labels are accessed.

E LIMITATIONS AND FUTURE WORK

The current implementation uses a fixed heterogeneous peer panel and therefore requires maintaining multiple VLMs at inference. Future work could investigate redundancy-aware peer selection and adaptive inference policies that reduce this cost while preserving the benefits of relational evidence.

Unlabeled answer relations and likelihood-derived confidence alone do not guarantee correct risk ordering in a new setting. Transfer depends on these signals remaining predictive of focal-answer correctness across settings.