Grounded or Guessing? LVLM Confidence Estimation via Blind-Image Contrastive Ranking

Anonymous ACL submission

Abstract

Large vision-language models suffer from visual ungroundedness: they can produce a fluent, confident, and even correct response driven entirely by language priors, with the image contributing nothing to the prediction. Existing confidence estimation methods cannot detect this, as they observe model behavior under normal inference with no mechanism to determine whether a prediction was shaped by the image or by text alone. We introduce **BICR** (**B**lind-**I**mage **C**ontrastive **R**anking), a model-agnostic confidence estimation framework that makes this contrast explicit during training by extracting hidden states from a frozen LVLM twice: once with the real image-question pair, and once with the image blacked out while the question is held fixed. A lightweight probe is trained on the real-image hidden state and regularized by a ranking loss that penalizes higher confidence on the blacked-out view, teaching it to treat visual grounding as a signal of reliability at zero additional inference cost. Evaluated across five modern LVLMs and seven baselines on a benchmark covering visual question answering, object hallucination detection, medical imaging, and financial document understanding, BICR achieves the best cross-LVLM average on both calibration and discrimination simultaneously, with statistically significant discrimination gains robust to cluster-aware analysis at 4–18× fewer parameters than the strongest probing baseline.[1]

1 Introduction

In Large Vision-Language Models (LVLMs), an answer can be confident and correct while being entirely driven by language priors. LVLMs function by prepending a sequence of visual tokens, produced by a vision encoder, to the language model's input context before any text generation begins. The language backbone receives both image and text as input, but *this does not mean the answer is driven by both*. Specific attention heads in LVLMs attend up to five times more strongly to text than to visual tokens (Yang et al., 2025), and answer distributions are nearly indistinguishable whether the image is provided or replaced with a blank input (Woo et al., 2025; Fu et al., 2025). A model can therefore produce a fluent, confident, and even correct response based almost entirely on learned linguistic priors, with the image contributing nothing to the prediction. This failure mode, which we term *visual ungroundedness*, is distinct from ordinary model error: the model is not reasoning incorrectly about the image, it is bypassing the image altogether.

Existing confidence estimation methods cannot tell visually ungrounded predictions apart from grounded ones. Hidden-state probing methods do carry useful discriminative signal in the LVLM setting (Li et al., 2024), but they share a common blind spot: a probe trained on representations extracted under normal inference sees only a single snapshot of the model's internal state, from which it cannot determine whether that state was shaped by the image or by text alone. Without exposure to that contrast during training, such a probe will assign indistinguishable confidence scores to grounded and ungrounded predictions, systematically overstating confidence in the latter.

Detecting visual ungroundedness still requires getting the two basic properties of a confidence score right, and doing so simultaneously remains an open challenge. A useful confidence estimator must satisfy two properties at once: *calibration* (expressed confidence matches empirical accuracy, so that a model assigning 80% confidence to a set of answers is correct on roughly 8 of 10 of them (Guo et al., 2017)), and *discrimination* (confidence scores meaningfully separate correct predictions from incorrect ones). LVLMs do not naturally

[1] **Anonymized code and sample data** are provided as separate Software and Data attachments in OpenReview.

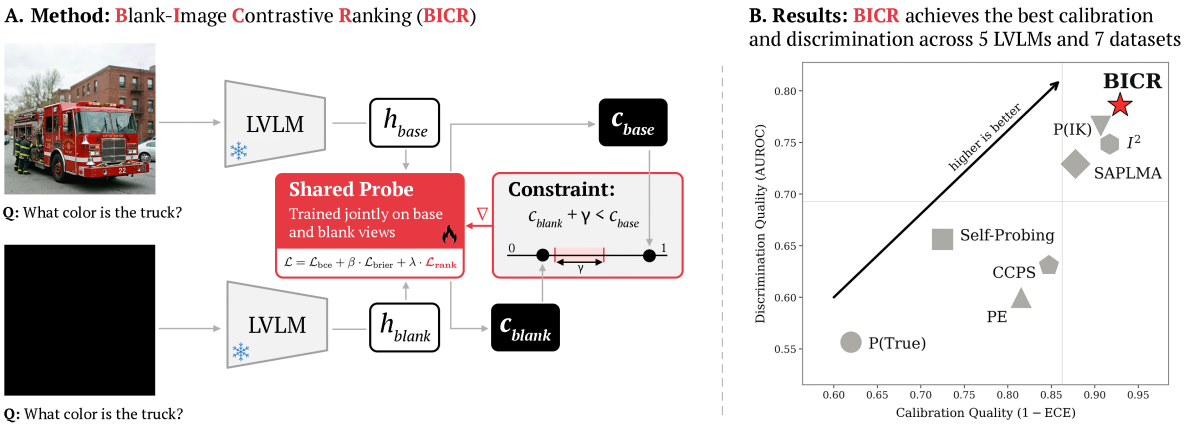


Figure 1: **Overview of our method (BICR) and headline results. (A)** BICR pairs each question with two views, the real image and a blank counterfactual, and trains a shared probe on top of a frozen large vision-language model (LVLM). The ranking loss $\mathcal{L}_{rank}$ enforces that the real-view confidence c_{base} exceeds the blank-view confidence c_{blank} by a margin γ, teaching the probe that confidence must be grounded in the visual input. Only the base view is used at inference. **(B)** Each marker is one confidence estimator's cross-LVLM average across our 5 LVLMs and 7 datasets, plotting calibration quality, $1 - \text{ECE}$ (Expected Calibration Error), against discrimination quality, AUROC (Area Under the ROC Curve). BICR (red star) is the only method in the upper-right region, achieving both high calibration and high discrimination simultaneously.

exhibit either: when asked a question they cannot reliably answer, LVLMs respond with the same fluent confidence as when they are correct, expressing no meaningful uncertainty (Guo et al., 2017; Dang et al., 2026; Chen et al., 2025). Prior methods have made progress on each property individually (see §2), but jointly satisfying both, especially under a visual modality, remains an open question.

Blind-Image Contrastive Ranking (BICR) addresses visual ungroundedness directly by teaching a confidence probe to distinguish image-driven predictions from language-prior-driven ones. For each sample, BICR extracts hidden states from a frozen LVLM twice, once with the original image-question pair and once with the image blacked out. A lightweight probe is trained on real-image hidden states with a ranking loss penalizing higher confidence on the blacked-out view, teaching visual grounding as a reliability signal at no inference cost. Across five LVLMs and seven baselines on a benchmark spanning visual question answering (VQA), object hallucination detection, medical imaging, and financial document understanding (Figure 1), BICR achieves the best cross-LVLM averages on calibration and discrimination simultaneously, with statistically significant discrimination gains robust to cluster-aware analysis and 4–18× fewer parameters than the strongest probing baseline, InternalInspector (Beigi et al., 2024). The contributions of this work are as follows:

- **BICR**: a model-agnostic confidence estimation framework for LVLMs that trains a lightweight probe via a blind-image contrastive ranking objective to use visual grounding as a reliability signal, achieving the best cross-LVLM average on calibration (ECE, Brier) and discrimination (AUCPR, AUROC).
- **VLCB**: a benchmark aggregating seven public VQA sources with model responses and correctness labels across five modern LVLMs, spanning general, medical, and financial reasoning domains, released publicly to support reproducibility and future work on confidence estimation in multimodal models.
- **Comprehensive evaluation** of seven confidence estimation baselines (P(True), Self-Probing, Prompt Ensemble, P(I Know), SAPLMA, InternalInspector, CCPS) across five LVLM architectures and diverse high-stakes domains, providing the first systematic benchmark of confidence estimation in LVLMs under a unified protocol that jointly reports calibration and discrimination.
- **Empirical evidence** that the gap between real and blank-image hidden states, our proxy for visual grounding, gives a reliable and discriminative signal of correctness across models, tasks, and domains.

2 Related Work

Confidence Estimation in Language Models. Confidence estimation methods for language models fall into four families. *Prompt-based* methods elicit a confidence signal through additional queries: verbalized confidence asks the model to state its certainty numerically (Xiong et al., 2024; Tian et al., 2023; Zeng et al., 2025), and self-consistency samples multiple responses and measures agreement (Wang et al., 2023). *Logit-based* methods read the output distribution directly, using token-level log-probabilities (Zhou et al., 2025) often rescaled via temperature scaling (Jiang et al., 2021; Guo et al., 2017). *Internal-state probing* methods train classifiers on hidden activations: P(IK) (Kadavath et al., 2022) probes the state before generation, SAPLMA (Azaria and Mitchell, 2023) reads hidden-layer activations on the generated response, and InternalInspector (Beigi et al., 2024) pools attention, feed-forward, and residual states across all layers through a learned encoder. *Internal-stability* methods probe representational response to perturbation: CCPS (Khanmohammadi et al., 2025) applies targeted adversarial perturbations to the final hidden state and trains a classifier on the resulting trajectory. We benchmark BICR against these seven methods in §5.

Calibration and Grounding in Vision-Language Models. LVLMs prepend visual tokens to the language model's input, but this architectural integration does not guarantee that visual content drives the answer (Yang et al., 2025; Woo et al., 2025; Fu et al., 2025). Even when output is correct, meaning-preserving image perturbations induce substantial shifts in internal representations (Wani et al., 2026), and LVLMs are persistently miscalibrated, with verbalized confidence poorly tracking correctness in ways prompting and temperature scaling do not correct (Dang et al., 2026; Chen et al., 2025; Xuan et al., 2025). Recent LVLM-specific methods address adjacent symptoms without producing a per-sample calibrated score: CSP (Zhao et al., 2025) and medical VQA frameworks (Du et al., 2025) target output-level calibration, while VL-Calibration (Xiao et al., 2026) decomposes confidence into visual and reasoning components via reinforcement learning fine-tuning. A parallel line targets hallucination rather than confidence: VCD (Leng et al., 2024) contrasts output distributions from original and noise-distorted images at every decoding step, and VL-Uncertainty (Zhang et al., 2024) clusters responses to perturbed inputs and reports hallucination-detection accuracy. None targets the joint calibration-and-discrimination protocol we adopt, and they collectively retrain the model, work at the output level, or add inference cost without addressing whether internal representations are anchored in visual evidence.

Hidden-State and Attention Probing in LVLMs. Hidden-state probing carries useful signal in LVLMs: classifiers on final-layer representations achieve strong hallucination detection across architectures and tasks (Li et al., 2024). Two adjacent lines read related internal signals for different purposes than per-sample confidence: SVAR (Jiang et al., 2025) uses the middle-layer visual attention ratio to classify object-level hallucination during generation, and TVI (Long et al., 2026) contrasts hidden states with and without the image to localize a visual integration point and quantify language prior, reported as a population-level Spearman correlation rather than a per-sample probability. However, when the visual input is degraded, probe performance degrades sharply with it (Li et al., 2024), revealing that these probes read the quality of the visual signal in the hidden state rather than whether any visual signal was used. A probe reading an ungrounded prediction and a grounded one may therefore see nearly identical representations, since neither was ever exposed to that contrast.

The missing contrastive signal. Across all four paradigms, no method trains a per-sample confidence estimator on the contrast between grounded and ungrounded representations. Probing methods (P(IK), SAPLMA, InternalInspector, SVAR) read snapshots under normal inference; output-level methods (CSP, VL-Calibration, VCD, VL-Uncertainty) do not expose internal grounding; TVI performs the contrast but as a population diagnostic, not a per-sample correctness-trained score. BICR fills this gap.

3 Vision-Language Model Confidence Estimation Benchmark (VLCB)

Evaluating confidence estimation methods requires a benchmark that provides correctness labels for model responses across diverse tasks and model families, supports a clear separation between training and test distributions, and spans the domains where reliable confidence scores matter most. To our knowledge, no existing resource satisfies all three, which is what VLCB is de-

Table 1: VLCB train, validation, and test splits.

Split	Source	N	%
Train	GQA (VQA)	20,000	100.0
Val	GQA (VQA)	5,000	100.0
Test	GQA (VQA)	12,568	41.2
	POPE (Object Hallucination)	9,000	29.5
	GMAI-MMBench (Medical Reasoning)	4,549	14.9
	MMMU-Pro 4-option (College Reasoning)	1,720	5.6
	MMMU-Pro 10-option (College Reasoning)	1,725	5.7
	MME-Finance (Chart Understanding)	892	2.9
	LLaVA-Wild (Visual Dialogue)	60	0.2
	Total	**30,514**	**100.0**

signed to address. We evaluate BICR alongside seven baselines spanning three paradigm families: prompt-based methods that treat the LVLM as a black box (P(True) (Kadavath et al., 2022), Self-Probing (Xiong et al., 2024), Prompt Ensemble (Zollo and Zemel, 2025)), internal-state probes that train lightweight classifiers on hidden-state snapshots (P(I Know) (Kadavath et al., 2022), SAPLMA (Azaria and Mitchell, 2023), InternalInspector (Beigi et al., 2024)), and one internal-stability method that reads representational robustness under perturbation (CCPS (Khanmohammadi et al., 2025)). Full baseline descriptions are provided in Appendix C.

Design principle. The central design choice in VLCB is deliberate distribution shift between training and evaluation. Confidence probe training uses GQA (Hudson and Manning, 2019) exclusively: 20,000 training and 5,000 validation samples stratified by question type, with short, unambiguous answers. The test set is detailed in Appendix B and combines a held-out GQA test split for in-distribution reference with six additional datasets unseen during training: POPE (Li et al., 2023) for object hallucination detection, GMAI-MMBench (Chen et al., 2024) for medical multimodal reasoning, MME-Finance (Gan et al., 2025) for financial chart understanding, MMMU-Pro (Yue et al., 2025) in 4-option and 10-option configurations for college-level reasoning, and LLaVA-in-the-Wild (Liu et al., 2023b) for Open-ended visual dialogue. Table 1 summarizes the splits. This out-of-domain setup makes performance meaningful: a confidence estimator that succeeds only on its training distribution offers no deployment guarantee.

Model coverage and response generation. We evaluate five open-weight instruction-tuned LVLMs: Qwen3-VL-8B, LLaVA-NeXT-13B, InternVL3.5-14B, Gemma-3-27B, and DeepSeek-VL2. Together they cover 4.5B to 27B active parameters, three distinct vision encoder lineages, and both dense and mixture-of-experts language model architectures. Full architectural specifications are provided in Appendix A. All five models are run on the complete train, validation, and test splits under identical generation conditions: greedy decoding with a maximum of 64 new tokens and images downscaled to a maximum of 2,048 pixels on the longer edge. Each model receives a semantically uniform prompt instructing it to answer briefly and completely; prompt delivery details vary by model chat template and are described in Appendix B.4.

Correctness labeling. Each generated response is assigned a binary correctness label by a `gpt-5-mini` judge applied uniformly across all seven source datasets and all five LVLMs. Using a single judge across task types, including multiple-choice questions where string matching would in principle suffice, ensures that formatting variation across LVLM chat templates does not introduce grading artifacts and that all correctness labels are produced by the same protocol. This practice is well established in the confidence estimation literature, and our approach follows the same protocols adopted by prior work (Kapoor et al., 2024; Khanmohammadi et al., 2025, 2026). Per-LVLM correctness statistics broken down by dataset and split are reported in Appendix B.

4 Blind-Image Contrastive Ranking (BICR)

Problem setup. Let $\mathcal{M}$ denote a frozen LVLM with hidden dimension d_h. Given a visual question (q, v) with question text q and image v, we perform a prompt-only forward pass through $\mathcal{M}$ and extract the hidden state at the last prompt token position, the point at which the model's representation of the full input is formed and generation would begin. We denote this hidden state $\mathbf{h}_{\text{base}} \in \mathbb{R}^{d_h}$. The goal is to learn a function $f : \mathbb{R}^{d_h} \to \mathbb{R}$ whose sigmoid $\sigma(f(\mathbf{h}_{\text{base}}))$ estimates the probability that the model's generated answer a is correct, using only $\mathbf{h}_{\text{base}}$ at inference time.

Two-view hidden state extraction. For each training sample, given question q, image v, and binary correctness label $y \in \{0, 1\}$ derived from the model's generated answer (with $y = 1$ denoting a correct response and $y = 0$ an incorrect one), we extract hidden states from $\mathcal{M}$ under two input conditions. The base view uses the original image:

$\mathbf{h}_{\text{base}}$ is the hidden state at the last prompt token of the forward pass over (q, v). The blank view substitutes a solid black image $v_\varnothing$ (RGB $(0, 0, 0)$, matching the spatial dimensions of v) while holding q fixed: $\mathbf{h}_{\text{blank}}$ is the hidden state at the same position from the forward pass over $(q, v_\varnothing)$. Both states come from the final decoder layer, so any difference between them is attributable solely to the visual input. The blank-image pass is computed once as a preprocessing step and reused across all training runs and seeds.

Probe architecture. The confidence probe f is a multi-layer perceptron with ReLU activations and dropout between layers, mapping $\mathbf{h} \in \mathbb{R}^{d_h}$ to a scalar logit. The same MLP is shared across both views during training: $\mathbf{h}_{\text{base}}$ and $\mathbf{h}_{\text{blank}}$ pass through f with identical parameters. This weight sharing is what allows the contrastive objective introduced below to shape the learned representation rather than merely fitting a separate decision threshold per view: a single probe must simultaneously produce high confidence for grounded real-image predictions and suppressed confidence for blank-image ones. The depth and width of the MLP are selected by Optuna hyperparameter search; full details are in Appendix F.

Training objective. The training loss combines three terms. The supervised loss is binary cross-entropy with positive-class weighting to handle label imbalance, where n_+ and n_- are the counts of correct and incorrect samples in training:

$$\mathcal{L}_{\text{bce}} = \text{BCE}(f(\mathbf{h}_{\text{base}}), y; w_+), \quad w_+ = n_-/n_+ \tag{1}$$

The calibration loss is a Brier score penalty on base-view predictions:

$$\mathcal{L}_{\text{brier}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i^{\text{base}} - y_i)^2, \quad \hat{p}_i^{\text{base}} = \sigma(f(\mathbf{h}_{\text{base},i})) \tag{2}$$

The visual grounding ranking loss is the core contribution of BICR. For each correctly answered sample ($y = 1$), the probe is required to assign higher confidence when the real image is present than when it is absent, with margin $\gamma > 0$ enforcing a minimum gap between the two scores rather than merely requiring one to exceed the other:

$$\mathcal{L}_{\text{rank}} = \frac{\sum_i \text{ReLU}\big(\gamma - (\hat{p}_i^{\text{base}} - \hat{p}_i^{\text{blank}})\big) y_i}{\sum_i y_i + \epsilon} \tag{3}$$

where $\hat{p}_i^{\text{blank}} = \sigma(f(\mathbf{h}_{\text{blank},i}))$ and $\epsilon = 10^{-8}$. The constraint is restricted to correct samples because the ranking direction has a clear semantic interpretation there: a correct answer that relied on the image should produce higher confidence with the image present than without it. For incorrect answers no directional constraint is warranted, as the answer is wrong regardless of visual grounding. The three terms are combined as:

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \beta\mathcal{L}_{\text{brier}} + \lambda\mathcal{L}_{\text{rank}} \tag{4}$$

where β, λ, and γ are selected by Optuna. The contribution of each term is validated empirically in Appendix H.1.

Inference. At test time, BICR requires a single prompt-only forward pass through the frozen LVLM to obtain $\mathbf{h}_{\text{base}}$, followed by a pass through the probe:

$$\hat{c} = \sigma\big(f(\mathbf{h}_{\text{base}})\big) \tag{5}$$

The blank-image pass plays no role at deployment, so BICR adds zero inference overhead relative to any single-view probe. Trainable parameter counts are compared in Appendix G.BICR

5 Results

All methods are evaluated on the VLCB test split across five LVLMs under a shared protocol. We report four primary metrics: Expected Calibration Error (ECE) and Brier Score (BS) for calibration, and Area Under the Precision–Recall Curve (AUCPR) and Area Under the ROC Curve (AUROC) for discrimination. Full definitions and additional metrics are in Appendix D. Trained methods report mean performance across five random seeds, each with 50 Optuna hyperparameter trials, selected via a composite validation score that jointly optimizes discrimination and calibration; the full training and selection protocol is described in Appendix E.

Main results. Table 2 presents the pooled aggregate performance of all methods across the five LVLMs. BICR achieves the best cross-LVLM average on all four metrics: ECE of 7.09%, BS of 18.41%, AUCPR of 87.47%, and AUROC of 78.63%. Among per-LVLM results, BICR leads on AUCPR and AUROC for every LVLM (five of five), and on ECE and BS for three of five LVLMs.

The two strongest baselines are P(I Know) and InternalInspector. P(I Know) is architecturally identical to BICR, an Optuna-tuned MLP over a single hidden-state vector, but is trained with BCE loss alone. BICR's improvement over P(I Know) (+2.1 AUROC, +1.2 AUCPR, −2.2 ECE on the cross-LVLM average) is therefore attributable to the training-time auxiliary losses (primarily the ranking signal from the blank-image comparison),

Table 2: Pooled aggregate performance across all LVLMs. Metrics reported as percentages (%). Arrows indicate direction of improvement. Best values per LVLM are **bolded**. Trained methods report mean across 5 seeds (50 Optuna trials each). Bottom-right: cross-VLM average. Extended per-dataset results in Appendix I.

DeepSeek-VL2					*Qwen3-VL-8B-Instruct*				
Method	***ECE*** ↓	***BS*** ↓	***AUCPR*** ↑	***AUROC*** ↑	***Method***	***ECE*** ↓	***BS*** ↓	***AUCPR*** ↑	***AUROC*** ↑
P(True)	34.07	37.43	68.09	52.66	P(True)	43.88	44.16	76.18	54.62
Self-Probing	35.07	37.35	74.19	62.22	Self-Probing	24.39	27.47	77.00	59.49
Prompt Ensembles	16.30	24.72	72.49	73.46	Prompt Ensembles	19.80	25.96	66.20	52.90
SAPLMA	12.83	21.51	79.69	77.14	SAPLMA	10.57	19.43	86.41	74.22
P(I Know)	8.50	19.26	84.58	78.49	P(I Know)	7.54	17.99	88.43	77.19
CCPS	7.70	22.14	76.97	70.94	CCPS	28.67	45.64	66.18	44.93
InternalInspector	7.39	18.95	84.54	79.31	InternalInspector	**5.43**	**16.95**	89.75	79.60
`BICR (Ours)`	**6.02**	**17.90**	**86.19**	**81.11**	`BICR (Ours)`	8.91	17.40	**90.26**	**80.14**
LLaVA-NeXT-Vicuna-13B					*InternVL3.5-14B*				
Method	***ECE*** ↓	***BS*** ↓	***AUCPR*** ↑	***AUROC*** ↑	***Method***	***ECE*** ↓	***BS*** ↓	***AUCPR*** ↑	***AUROC*** ↑
P(True)	26.23	30.93	67.91	54.73	P(True)	41.19	41.45	78.02	59.49
Self-Probing	28.92	29.97	81.87	67.30	Self-Probing	21.51	24.71	76.77	70.75
Prompt Ensembles	12.77	23.49	77.31	68.47	Prompt Ensembles	16.70	25.90	60.26	43.37
SAPLMA	16.52	22.99	82.01	72.47	SAPLMA	16.58	23.27	76.21	65.42
P(I Know)	10.81	19.87	87.06	77.32	P(I Know)	10.83	20.16	86.35	73.78
CCPS	16.27	21.98	76.00	72.89	CCPS	14.72	23.83	71.64	58.22
InternalInspector	13.79	22.31	81.42	71.87	InternalInspector	10.59	21.48	82.64	69.31
`BICR (Ours)`	**5.65**	**18.16**	**87.74**	**78.94**	`BICR (Ours)`	**7.90**	**19.04**	**88.04**	**76.39**
Gemma-3-27B-IT					***Cross-LVLM Average***				
Method	***ECE*** ↓	***BS*** ↓	***AUCPR*** ↑	***AUROC*** ↑	***Method***	***ECE*** ↓	***BS*** ↓	***AUCPR*** ↑	***AUROC*** ↑
P(True)	44.80	45.05	74.01	56.81	P(True)	38.04	39.80	72.84	55.66
Self-Probing	27.67	29.46	79.20	68.39	Self-Probing	27.51	29.79	77.81	65.63
Prompt Ensembles	26.72	30.29	70.13	61.82	Prompt Ensembles	18.46	26.07	69.28	60.00
SAPLMA	4.61	**19.42**	83.83	75.45	SAPLMA	12.22	21.32	81.63	72.94
P(I Know)	8.88	19.86	85.07	76.08	P(I Know)	9.31	19.43	86.30	76.57
CCPS	8.98	22.14	73.37	68.49	CCPS	15.27	27.15	72.83	63.10
InternalInspector	**4.29**	19.68	83.17	74.15	InternalInspector	8.30	19.88	84.31	74.85
`BICR (Ours)`	6.98	19.56	**85.10**	**76.56**	`BICR (Ours)`	**7.09**	**18.41**	**87.47**	**78.63**

Table 3: Loss ablation for `BICR` (cross-LVLM average, 5 seeds × 5 LVLMs).

Variant	**ECE**↓	**BS**↓	**AUCPR**↑	**AUROC**↑
Full (`BICR`)	**7.09**	**18.41**	**87.47**	**78.63**
$-\mathcal{L}_{\text{brier}}$	8.48 (+1.39)	19.04 (+0.63)	87.10 (-0.37)	78.02 (-0.61)
$-\mathcal{L}_{\text{rank}}$	8.13 (+1.04)	19.63 (+1.22)	85.52 (-1.95)	75.31 (-3.32)
$\mathcal{L}_{\text{bce}}$ only	9.15 (+2.06)	19.91 (+1.50)	85.48 (-1.99)	75.25 (-3.38)

not to architectural differences. InternalInspector achieves competitive ECE on two LVLMs (Qwen and Gemma) using a ResNet18-based CNN encoder with 11.3M parameters, 7× more than `BICR`'s average of 1.6M, but falls behind on discrimination across all LVLMs. A detailed comparison of trainable parameter counts is provided in Appendix G. The three prompt-based methods (P(True), Self-Probing, Prompt Ensemble) trail the internal-state probes by sizable margins on the cross-LVLM average, with the smallest gap (Self-Probing to SAPLMA) at roughly 7 AUROC points and the largest exceeding 20, confirming that verbalized or logit-based confidence signals are insufficient for reliable confidence estimation without access to internal representations. All discrimination improvements of `BICR` over trained baselines are statistically significant under both pooled and cluster-aware analyses, and calibration improvements are significant against three of the four trained baselines (P(I Know), SAPLMA, CCPS) under both analyses; full significance results are reported in Appendix I.8.

Ablation study. Table 3 reports the contribution of each loss component to `BICR`'s training objective. $\mathcal{L}_{\text{rank}}$ is the critical component: removing it degrades AUROC by 3.32 points and AUCPR by 1.95 ($p < 0.001$, paired Wilcoxon and cluster-aware bootstrap), confirming that the blank-image contrastive signal is the primary driver of `BICR`'s discriminative gain. $\mathcal{L}_{\text{brier}}$ provides a smaller but consistent calibration benefit ($\Delta\text{ECE} = -1.39$, $\Delta\text{BS} = -0.63$). Removing both auxiliary losses produces the worst configuration on every metric

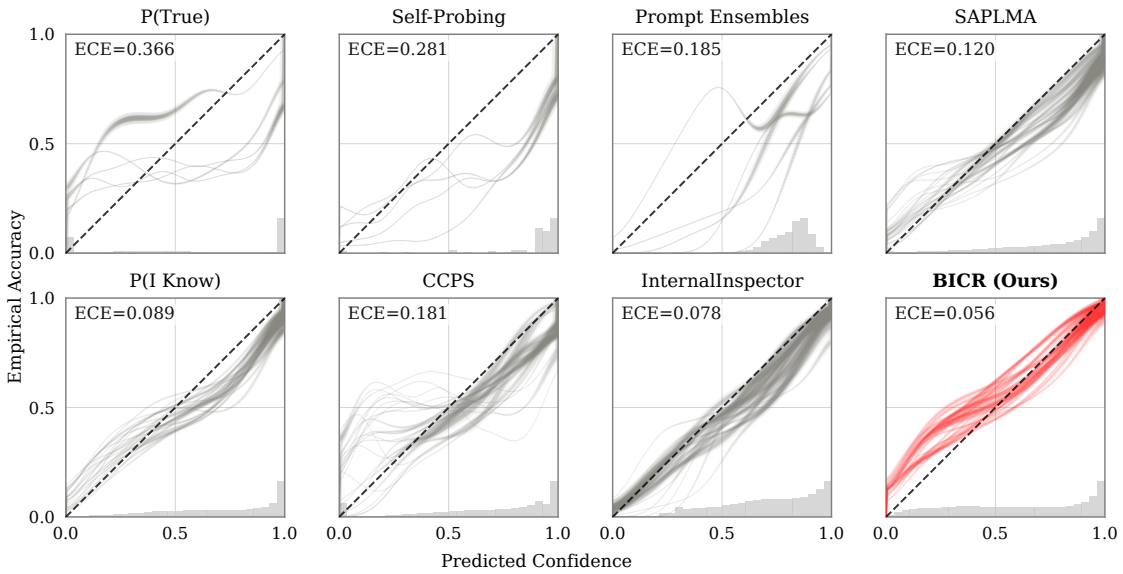


Figure 2: Cross-LVLM reliability diagrams for all eight methods. ECE (top-left of each panel) is computed on samples pooled across all 5 LVLMs and 5 seeds, matching the reliability curves; these values differ slightly from the per-LVLM-then-averaged ECE in Table 2 because ECE is non-linear in the per-bin frequencies. Bar height is the empirical accuracy for that confidence bin, bar opacity is proportional to the fraction of samples in the bin, and the dashed diagonal represents perfect calibration.

($p < 0.005$ on all discrimination metrics). Beyond the loss components, three additional design studies are reported in the appendix: a behavioral analysis showing that $\mathcal{L}_{\text{rank}}$ broadens the probe's confidence distribution and tightens calibration in the high-confidence range where overconfidence is most consequential (Appendix H.2); a comparison of five null-image strategies (black, white, Gaussian noise, blurred original, pixel-shuffled), in which the solid-black null is the strongest training signal on every metric (Appendix H.4); and an analysis of the Optuna-selected loss coefficients, which finds the rank weight λ consistently positive and non-trivial across LVLMs and seeds (Appendix H.3). Per-LVLM ablation breakdowns and full significance results are in Appendix H.1.

Calibration analysis. Figure 2 presents the cross-LVLM reliability diagram for all eight methods. Three patterns are evident. Prompt-based methods exhibit severe miscalibration: P(True) concentrates most of its mass in the extreme bins ($[0, 0.2)$ and $[0.8, 1.0]$) with empirical accuracies far from the predicted values (ECE $= 0.366$); Self-Probing pushes nearly all predictions above 0.8 regardless of correctness (ECE $= 0.281$). Trained baselines improve substantially, with InternalInspector reaching ECE $= 0.078$, but concentrate predictions in the mid-to-high range, underutilizing the low-confidence region where uncertain predictions should fall. BICR achieves the best calibration (ECE $= 0.056$) with confidence mass distributed across the full range: the $[0.8, 1.0]$ bin achieves 91.3% empirical accuracy, and substantial mass at $[0.2, 0.4)$ reflects 42.1% accuracy, indicating that BICR has learned to express genuine uncertainty when visual evidence is weak. A bin-level analysis showing that $\mathcal{L}_{\text{rank}}$ is responsible for this behavior is provided in Appendix H.2.

Per-dataset analysis. BICR achieves the best pooled performance across all datasets combined (Table 2), and this advantage holds under per-dataset equal-weight aggregation as well: averaging metrics with equal weight per dataset (rather than per sample) widens BICR's ECE and BS gaps over every trained baseline (Appendix I.3), confirming that the headline result is not an artifact of the larger datasets dominating the pooled view. The per-dataset breakdown (Appendix I) further shows that BICR's gains decompose into two regimes. On the larger datasets (GQA, POPE), BICR's discrimination advantage is consistent across LVLMs, sharpening the correct–incorrect ranking on tasks where representations are already informative. On the harder grounding-bound datasets (GMAI-MMBench, MMMU-Pro, MME-Finance), discrimination is more contested, but BICR's calibration advantage is largest there: it is the best-calibrated method on four of seven datasets (§I.6), reflecting that the blank-image contrastive signal prevents the probe from collapsing into overconfident outputs

in regimes where every method struggles to reason about the image. This split between where discrimination concentrates (large, easier datasets) and where calibration concentrates (harder grounding-bound datasets) is what produces BICR's joint advantage on both axes when aggregated across the benchmark. A direct mechanism check on the sub-population the LVLM behaviorally treats as image-invariant (Appendix J) confirms that BICR's confidence is the most accurate of any method evaluated on samples where the image is genuinely not driving the answer, with paired-bootstrap BS gains significant on 31 of 35 (LVLM, baseline) pairs.

6 Discussion

Visual grounding is readable from hidden states, and its absence is detectable. The gap between hidden states under a real versus blank image is a reliable signal of correctness, even though a probe trained only on real-image states has no basis to distinguish a grounded prediction from one shaped entirely by language priors. The 3.32 AUROC drop when $\mathcal{L}_{\text{rank}}$ is removed ($p < 0.001$ across 25 runs and all five LVLMs) confirms the signal is not incidental, and a behavioral test on the sub-population the LVLM treats as image-invariant (Appendix J) shows BICR's calibration advantage concentrates on exactly the population the rank loss targets.

Confidence estimation can precede response generation entirely. BICR operates at the last prompt token, after the model has formed its full internal representation of the task but before generation begins. Confidence is therefore estimated from how the model represented the input, not from what it said. In deployment, a pipeline can flag low-confidence inputs before paying generation cost, enabling triage, escalation, or human review without waiting for a response. P(True) and Self-Probing require the model to generate an answer and then a second response expressing certainty; SAPLMA and CCPS are bound to a completed generation pass. BICR has no such dependency.

Training-time contrast, no inference-time cost. The contrastive signal that drives BICR's gains could in principle be read at inference by running both passes per query. BICR instead absorbs the contrast into the probe's weights during training: the blank-image pass is preprocessing, and inference is a single forward pass. Methods like Prompt Ensemble, Self-Probing, and CCPS pay their cost at every query; BICR pays it once.

The calibration-discrimination trade-off yields to the right training signal. Prior work identifies a tension where optimizing one metric degrades the other (Khanmohammadi et al., 2026). BICR is the only method in our benchmark with the best cross-LVLM average on all four metrics simultaneously, and the ablation reveals why: $\mathcal{L}_{\text{rank}}$ improves discrimination by suppressing overconfident scores on ungrounded predictions, while $\mathcal{L}_{\text{brier}}$ penalizes the gap between predicted scores and true correctness frequencies. The two losses are complementary because they target distinct failure modes: ranking correct above incorrect, versus producing scores reliable as probability estimates.

BICR's advantage concentrates where it matters most. Gains are largest on GMAI-MMBench and MMMU-Pro, the two hardest and most visually demanding benchmarks in VLCB, and smallest on POPE, where binary hallucination detection is learnable largely from language priors. This validates the design: the contrastive signal is most informative where the image is genuinely necessary. An ungrounded correct guess on a yes/no probe is an edge case; on a medical diagnosis or financial chart, the same failure carries real cost.

7 Conclusion

We introduced BICR, an LVLM confidence estimation framework built on a targeted training-time intervention: replace the image with black, extract the hidden state at the same prompt position, and regularize a lightweight probe to assign lower confidence when the visual input disappears. This single contrastive signal, applied before generation begins at zero inference overhead, is sufficient to achieve state-of-the-art performance on calibration and discrimination simultaneously across five LVLMs and seven baselines, with statistically significant gains at 4–18$\times$ fewer parameters than the strongest probing baseline. The findings suggest that grounding is legible in hidden states, that the calibration-discrimination trade-off yields to the right training signal, and that higher confidence quality does not require more inference compute. We release VLCB, our LVLM confidence estimation benchmark spanning general, medical, and financial visual reasoning, together with all evaluation code to support future research on trustworthy LVLM deployment.

Limitations

Despite BICR's strong empirical performance across five LVLMs and seven VQA datasets, several limitations bound the scope of our findings and point to natural directions for future work. First, BICR requires access to the LVLM's internal hidden states, which is feasible for open-weight models but precludes deployment on closed-weight LVLMs (e.g., GPT-4V, Claude, Gemini) accessed only through APIs; extending the blank-image contrastive principle to API-only settings is an open problem. Second, we do not benchmark against finetuning-based methods such as calibration-tuning (Kapoor et al., 2024), which retrain the base LVLM; the cost of finetuning each of our five LVLMs across seeds and a hyperparameter search would have been prohibitive at our benchmark's scale, and a fair comparison would require its own dedicated study. Third, VLCB is dominated in absolute sample count by GQA and POPE; while we report both pooled and equal-weight aggregations to surface this asymmetry, the dataset mix is not exhaustive of all visual reasoning regimes (notably absent: video, 3D, and embodied settings). Fourth, our correctness annotations rely on an LLM judge (GPT-5-mini) rather than expert human annotation; while LLM-as-judge is standard at this scale (approximately 150,000 graded responses), expert grading of the medical imaging and document understanding subsets would be a valuable validation step. Fifth, BICR's rank loss suppresses confidence on correct-but-ungrounded predictions, which AUROC and ECE penalize. We accept this trade-off: such predictions are correct by accident, and suppressing confidence in deployment is the desired behavior even if a one-shot benchmark counts it against the estimator. The effect is empirically small, since correct-but-ungrounded predictions are a minority on the visually demanding datasets where BICR's gains concentrate. Sixth, even with these gains, BICR does not achieve perfect calibration: on the hardest reasoning datasets (GMAI-MMBench, MMMU_Pro), every method we evaluate is systematically overconfident in the mid-to-high confidence range, and BICR, while consistently the closest to the diagonal in those panels, does not fully close the gap. The blank-image contrast targets visual ungroundedness and is therefore best positioned to correct errors at the visual integration stage; errors that arise downstream, when the model uses the image and still reasons incorrectly about it, leave real-image and blank-image hidden states similarly grounded and offer the rank loss little signal to act on. Closing this remaining gap is a separate problem from the one this work addresses. Seventh, our use of "visual grounding" and "visual ungroundedness" refers to an operational proxy rather than a direct measurement: we treat a prediction as grounded to the extent that the LVLM's hidden state differs between the real-image and blank-image views, a necessary but not sufficient condition for grounding in the stronger semantic sense. This contrastive-proxy framing is shared with the broader LVLM grounding literature, including VCD (Leng et al., 2024), VL-Uncertainty (Zhang et al., 2024), and SVAR (Jiang et al., 2025), none of which directly measure grounding either; more direct measurements (e.g., causal interventions on individual visual tokens) are an open problem. Eighth, the rank loss $\mathcal{L}_{\text{rank}}$ applies a directional constraint only to correctly-answered samples ($y = 1$) and is unweighted on the incorrect class. The asymmetry follows from the design intent (the contrastive direction encodes *grounded correctness*, which has no analogue for incorrect responses where the answer is wrong regardless of grounding), but it leaves the incorrect-but-ungrounded versus incorrect-but-grounded distinction unexploited at training time; a symmetric or class-conditional formulation that also constrains the $y = 0$ direction is a natural extension we did not investigate. Finally, the five LVLMs we evaluate range from 8B to 27B parameters and are all open-weight English-language instruction-tuned models; results may not transfer cleanly to substantially smaller or larger models, to multilingual settings, or to LVLMs trained primarily for non-VQA objectives. We believe these limitations do not detract from our core findings but instead provide a clear roadmap for future investigations into the reliability of LVLMs.

Ethical Considerations

While BICR is developed with the goal of improving the reliability of large vision-language models, several ethical considerations are relevant. The primary concern is over-reliance on automated confidence scores. Our results show that even the best methods carry trade-offs and no method is perfectly calibrated across every dataset and LVLM combination. In high-stakes domains such as medicine, finance, or law, accepting an LVLM's output sim-

ply because its associated confidence score is high, without independent human judgment and oversight, could lead to adverse outcomes when an LVLM error is not flagged by the confidence estimator. This concern is especially acute in the medical imaging and document understanding settings represented in our benchmark, where confidence scores might inform downstream decisions with material consequences. A second concern is fairness across diverse populations and data distributions. The LVLMs we evaluate carry whatever biases were present in their training data, and any confidence estimator built on top of those LVLMs, including BICR, may inherit or even amplify those biases. As a result, confidence scores could be systematically less reliable for certain demographic groups, image styles, or question types, potentially leading to inequitable downstream outcomes. Therefore, any deployment of BICR or related methods, particularly in sensitive applications, should be preceded by thorough fairness testing across relevant subgroups, accompanied by ongoing monitoring, and framed explicitly as a tool that assists human experts rather than replacing their critical judgment.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *Preprint*, arXiv:2304.13734.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. 2024. InternalInspector i^2: Robust confidence estimation in LLMs through internal states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12847–12865, Miami, Florida, USA. Association for Computational Linguistics.

Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. 2024. Finetuning language models to emit linguistic expressions of uncertainty. *Preprint*, arXiv:2409.12180.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Yu Qiao, and Junjun He. 2024. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, Zheng Zhang, and Richang Hong. 2025. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3095–3109, Abu Dhabi, UAE. Association for Computational Linguistics.

Yunkai Dang, Yifan Jiang, Yizhu Jiang, Anqi Chen, Wenbin Li, and Yang Gao. 2026. Instinct vs. reflection: Unifying token and verbalized confidence in multimodal large models. *Preprint*, arXiv:2604.17274.

Yuetian Du, Yucheng Wang, Ming Kong, Tian Liang, Qiang Long, Bingdi Chen, and Qiang Zhu. 2025. Confidence calibration for multimodal LLMs: An empirical study through medical VQA. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part VI*, page 89–99, Berlin, Heidelberg. Springer-Verlag.

Stephanie Fu, tyler bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*.

Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. 2025. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 12867–12874, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *Preprint*, arXiv:2411.16724.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14, St Julians, Malta. Association for Computational Linguistics.

Reza Khanmohammadi, Erfan Miahi, Simerjot Kaur, Charese Smiley, Ivan Brugere, Kundan S Thind, and Mohammad M. Ghassemi. 2026. How reliable are confidence estimators for large reasoning models? a systematic benchmark on high-stakes domains. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1754, Rabat, Morocco. Association for Computational Linguistics.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese Smiley, Kundan S Thind, and Mohammad M. Ghassemi. 2025. Calibrating LLM confidence by probing perturbed representation stability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10448–10514, Suzhou, China. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.

Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, and Fakhri Karray. 2024. Reference-free hallucination detection for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4542–4551, Miami, Florida, USA. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. 2026. Understanding language prior of lvlms by contrasting chain-of-embedding. *Preprint*, arXiv:2509.23050.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. *Preprint*, arXiv:2002.07650.

Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. 2025. Trained on tokens, calibrated on concepts: The emergence of semantic calibration in llms. *Preprint*, arXiv:2511.04869.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Farooq Ahmad Wani, Alessandro Suglia, Rohit Saxena, Aryo Pradipta Gema, Wai-Chung Kwan, Fazl Barez, Maria Sofia Bucarelli, Fabrizio Silvestri, and Pasquale Minervini. 2026. Same answer, different representations: Hidden instability in vlms. *Preprint*, arXiv:2602.06652.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1927–1951, Vienna, Austria. Association for Computational Linguistics.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Wenyi Xiao, Xinchi Xu, and Leilei Gan. 2026. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. *Preprint*, arXiv:2604.09529.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. 2025. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1450, Suzhou, China. Association for Computational Linguistics.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.

Qingcheng Zeng, Weihao Xuan, Leyang Cui, and Rob Voigt. 2025. Thinking out loud: Do reasoning models know when they're right? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1407, Suzhou, China. Association for Computational Linguistics.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. *Preprint*, arXiv:2504.05419.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2024. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *Preprint*, arXiv:2411.11919.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Ruibo Hou, Jiaming Guo, Zihao Zhang, Yifan Hao, and Yunji Chen. 2025. Object-level verbalized confidence calibration in vision-language models via semantic perturbation. *Preprint*, arXiv:2504.14848.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. 2025. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*.

Thomas P Zollo and Richard Zemel. 2025. Confidence calibration in vision-language-action models. *Preprint*, arXiv:2507.17383.

Appendix

Table of Contents

A Large Vision-Language Model Backbones **15**

B VLCB Benchmark Construction **15**
- B.1 Data Curation and Standard Schema . . . 15
- B.2 Source Datasets and Versions . . . 16
  - B.2.1 GQA (Compositional Visual Reasoning) . . . 17
  - B.2.2 POPE (Object Hallucination Probing) . . . 17
  - B.2.3 GMAI-MMBench (Medical Multimodal Reasoning) . . . 17
  - B.2.4 MME-Finance (Financial Chart VQA) . . . 18
  - B.2.5 MMMU-Pro (Multidisciplinary College-Level VQA) . . . 18
  - B.2.6 LLaVA-in-the-Wild (Open-Ended Multimodal Instruction Following) . . . 18
- B.3 Final Train, Validation, and Test Assembly . . . 18
- B.4 Response Generation . . . 19
- B.5 Response Grading . . . 20
- B.6 Benchmark Availability and Licensing . . . 25

C Baseline Confidence Estimation Methods **25**
- C.1 Prompt-Based Methods . . . 25
  - C.1.1 P(True) . . . 25
  - C.1.2 Self-Probing (SP) . . . 26
  - C.1.3 Prompt Ensembles (PE) . . . 26
- C.2 Internal-State Probing Methods . . . 27
  - C.2.1 P(I Know) (P(IK)) . . . 27
  - C.2.2 SAPLMA . . . 27
  - C.2.3 InternalInspector (I^2) . . . 28
- C.3 Internal-Stability Methods . . . 28
  - C.3.1 CCPS . . . 28

D Evaluation Metrics **29**
- D.1 Calibration Metrics . . . 29
  - D.1.1 Expected Calibration Error (ECE) . . . 29
  - D.1.2 Brier Score (BS) . . . 29
- D.2 Discrimination Metrics . . . 29
  - D.2.1 Accuracy (ACC) . . . 29
  - D.2.2 F1 Score (F1) . . . 29
  - D.2.3 Area Under the Precision–Recall Curve (AUCPR) . . . 29
  - D.2.4 Area Under the ROC Curve (AUROC) . . . 30

E Validation Monitoring and Model Selection **30**

F Hyperparameter Search with Optuna **30**
- F.1 Optimization Protocol . . . 30
- F.2 Hyperparameter Search Space . . . 31

G Analysis of Additional Trainable Parameters **31**
G.1 Base LVLM Architectural Parameters . . . 31
G.2 Formulation of Additional Trainable Parameters . . . 31
G.3 Exact Additional Trainable Parameter Counts . . . 32
G.4 Discussion . . . 32

H Design Validation and Ablation Analysis **33**
H.1 Loss Component Ablation . . . 33
H.2 What $\mathcal{L}_{\text{rank}}$ Teaches the Probe . . . 34
H.3 Optimized Hyperparameter Analysis . . . 35
H.4 Choice of Null Image . . . 35

I Extended Results **37**
I.1 Per-LVLM Pooled Performance . . . 37
I.2 Cross-LVLM Pooled Average . . . 44
I.3 Per-LVLM Unweighted Average Across Datasets . . . 44
I.4 Cross-LVLM Unweighted Average . . . 48
I.5 Per-LVLM Per-Dataset Breakdown . . . 48
I.6 Per-LVLM and Per-Dataset Reliability Diagrams . . . 56
I.7 Loss Component Ablation . . . 65
I.8 Statistical Significance . . . 65

J Direct Behavioral Test: Calibration on Image-Invariant Samples **66**
J.1 Sub-population: Behaviorally Image-Invariant Samples . . . 66
J.2 Methodology . . . 67
J.3 Headline Result on the Image-Invariant Subset . . . 68
J.4 Failure Subset (Image-Invariant, Confidently Wrong) . . . 68
J.5 Cross-Subset Win Summary . . . 69
J.6 Discussion . . . 70

A Large Vision-Language Model Backbones

The five LVLMs benchmarked in this work were selected to make our cross-model claims meaningful rather than to maximize coverage of any single architectural axis. Together they span four open-weight model families with dense language backbones (Qwen3-VL, LLaVA-NeXT, InternVL3.5, Gemma-3) and one mixture-of-experts design (DeepSeek-VL2); a parameter range from 4.5B activated to 27B; three distinct vision-encoder lineages (CLIP-derived ViT-L/14, the SigLIP family, and the Qwen3-VL ViT with DeepStack); and language context lengths from 4K to 256K tokens. The intent of this selection is that any confidence estimation method evaluated across all five must demonstrate it generalizes across visual encoders, language backbones, and parameter regimes rather than exploiting properties of a particular architecture.

Tables 4 and 5 report the LLM-side and vision-encoder configurations respectively, taken directly from the official model configs on the Hugging Face Hub at the time of writing. Two structural points in these tables are worth noting because they shape how BICR's hidden-state extraction (§4) interacts with each model. First, hidden size H varies from 2,560 (DeepSeek-VL2) to 5,376 (Gemma-3), which directly drives the trainable parameter count of every probe-based confidence estimator we evaluate; the exact per-model parameter counts are reported in Appendix G. Second, the vision encoders differ not only in lineage but in input resolution policy: Qwen3-VL accepts dynamic resolutions, SigLIP variants are fixed at 384 or 896 pixels, and CLIP ViT-L/14 is fixed at 336 pixels, which is what motivates the uniform 2,048-pixel cap applied at the input side of our generation pipeline (§B.4) before each model's own preprocessor takes over.

Compute environment. The pipeline uses a two-tier setup that separates expensive LVLM inference from lightweight confidence-estimator training. LVLM inference — hidden-state extraction, response generation for VLCB construction, and the inference-only baselines (Self-Probing, Prompt Ensembles) — runs on NVIDIA H200 GPU, with response generation served through vLLM (Kwon et al., 2023) for high throughput. Four of the five LVLMs (Qwen3-VL-8B, LLaVA-NeXT-13B, InternVL3.5-14B, Gemma-3-27B) run in full precision; DeepSeek-VL2 runs in half precision with reduced batch size due to known numerical instabilities (precision settings are detailed in Appendix B). Confidence-estimator training and evaluation (BICR, P(I Know), SAPLMA, InternalInspector, CCPS) operate on cached hidden states rather than the live LVLM and run on a cluster of 8×NVIDIA A100 40GB GPUs. Each (method, LVLM, seed) configuration fits on a single A100, so we shard the (LVLM, seed) grid across the 8 GPUs — the 50 Optuna trials per (LVLM, seed) tuple for BICR and P(I Know) are the dominant training cost.

B VLCB Benchmark Construction

The main body (§3) frames VLCB by its design principle: training and validation are drawn from a single source on purpose, and evaluation spans heterogeneous out-of-domain task formats on purpose, so that performance numbers reflect generalization rather than within-distribution fitting. This appendix documents the engineering that operationalizes that principle. We describe the seven public source datasets aggregated into VLCB, the unified record schema that absorbs their heterogeneity into a single training and evaluation interface, the train, validation, and test assembly together with the quality-control checks that guarantee its splits are disjoint and reproducible, the response generation pipeline used to elicit answers from each evaluated LVLM under semantically uniform conditions, and the LLM-judge protocol used to label every generated response as correct or incorrect. All datasets are in English. For comprehensive information regarding the original construction and domain coverage of each source benchmark, we refer the reader to their respective publications.

B.1 Data Curation and Standard Schema

The seven source datasets aggregated into VLCB span grounded visual reasoning, hallucination probing, multimodal multiple-choice exams, medical VQA, financial chart understanding, and open-ended instruction following. Each was originally released with its own record format, image storage convention, and label schema, none of which agree across sources. To support a unified training and evaluation harness across this heterogeneity, every raw sample is processed into the following standardized HuggingFace Dataset record:

- **hash_id** (str): a deterministic, dataset-specific MD5 hash over {dataset}[SEP]

Table 4: LLM-side architecture of the five LVLMs benchmarked in VLCB. Columns report the Hugging Face model identifier (Hub ID), language backbone (Backbone), parameter count in billions (P, with activated parameters reported for DeepSeek-VL2), hidden size (H), number of transformer layers (L), number of attention heads as query/key-value heads ($H\,(Q/KV)$), and maximum context length (C). Numbers are taken from official model configs on the Hugging Face Hub at the time of writing. †DeepSeek-VL2 uses an MoE LM; activated parameters are reported.

Hub ID	**Backbone**	P	H	L	$H\,(Q/KV)$	C
Qwen/Qwen3-VL-8B-Instruct (Team, 2025)	Qwen3-VL-Text	8.0	4096	36	32 / 8	256K
llava-hf/llava-v1.6-vicuna-13b-hf (Liu et al., 2023a)	LLaMA (Vicuna-13B)	13.0	5120	40	40 / 40	4K
OpenGVLab/InternVL3_5-14B-HF (Wang et al., 2025)	Qwen3-14B	15.1	5120	40	40 / 8	40K
google/gemma-3-27b-it (Gemma Team et al., 2025)	Gemma-3	27.0	5376	62	32 / 16	128K
deepseek-ai/deepseek-vl2 (Wu et al., 2024)	DeepSeek-VL2 (MoE)†	4.5	2560	30	32 / 32	4K

Table 5: Vision-encoder architecture of the five LVLMs benchmarked in VLCB. Numbers are taken from official model configs on the Hugging Face Hub at the time of writing.

Hub ID	**Vision Encoder**	**ViT Hidden**	**ViT Layers**	**ViT Heads**	**Patch**	**Input Res.**
Qwen/Qwen3-VL-8B-Instruct	Qwen3-VL ViT	1152	27	16	16	dynamic
llava-hf/llava-v1.6-vicuna-13b-hf	CLIP ViT-L/14	1024	24	16	14	336
OpenGVLab/InternVL3_5-14B-HF	InternViT-300M	1024	24	16	14	448
google/gemma-3-27b-it	SigLIP	1152	27	16	14	896
deepseek-ai/deepseek-vl2	SigLIP-SO400M	1152	27	16	14	384

`{category}[SEP]{question}[SEP]{answer}[SEP]{image_key}`, where `image_key` is whichever per-source identifier (image filename, base64 payload, image-slot list) uniquely disambiguates samples that would otherwise share question and answer text.

- **`image`** (`PIL.Image.Image`, RGB): the visual input. Variable resolution is preserved at curation time; resizing is applied only at inference (§B.4).
- **`question`** (`str`): the input question presented to the LVLM, with task-specific multiple-choice options and any required context already inlined.
- **`answer`** (`str`): the ground-truth answer in the canonical form expected for that source dataset (e.g., a single option letter for multiple-choice tasks; a short answer span for GQA; the textual answer for MME-Finance and LLaVA-in-the-Wild).
- **`category`** (`str`): a dataset-specific sub-category label (e.g., GQA's `detailed` question type, POPE's negative-sampling regime, GMAI-MMBench's clinical VQA task, MMMU-Pro's subject and topic difficulty). Defaults to `"N/A"` when no taxonomy is available.
- **`dataset`** (`str`): the source dataset identifier.

The `hash_id` field is the mechanism that makes VLCB reproducible by independent users. Because the licensing terms of several source datasets prevent us from redistributing the assembled benchmark as a single archive (§B.6), reproducibility relies on the property that any user who has independently obtained the source datasets can recompute the same hashes and recover the same splits. To support this, all curation procedures fix the random seed at `SEED=23` so that subset selection, stratified splitting, and ordering are fully deterministic. Figure 3 shows one representative sample from each source dataset to illustrate the visual and textual variety the schema absorbs.

B.2 Source Datasets and Versions

Each source dataset is loaded from its official distribution and processed independently into a per-source artifact before any cross-source aggregation. Table 6 lists every source, its access path, the split used, and its original license.

Table 6: Source datasets used to construct VLCB, with split selection and license.

Dataset	Source	Split / License
GQA (Hudson and Manning, 2019)	HuggingFace lmms-lab/GQA	balanced train, val, testdev / MIT
POPE (Li et al., 2023)	HuggingFace lmms-lab/POPE	test (adversarial / popular / random) / CC BY 4.0
GMAI-MMBench (Chen et al., 2024)	HuggingFace OpenGVLab/GMAI-MMBench	VAL TSV (only split with answers) / Apache 2.0
MME-Finance (Gan et al., 2025)	Official MME-Finance release (TSV + image archive)	CC BY-NC-SA 4.0
MMMU-Pro (Yue et al., 2025)	HuggingFace MMMU/MMMU_Pro	standard (4 options), standard (10 options) / Apache 2.0
LLaVA-in-the-Wild (Liu et al., 2023b)	Official LLaVA-Bench-in-the-Wild release	test / -

B.2.1 GQA (Compositional Visual Reasoning)

GQA (Hudson and Manning, 2019) is a large-scale benchmark for compositional visual question answering over real-world scene graphs, and is the source of every training and validation sample in VLCB; the held-out testdev split appears in the test set as the in-distribution reference point. We use the *balanced* train, val, and testdev splits released through lmms-lab/GQA, which contain 943,000 / 132,062 / 12,578 instructions over 72,140 / 10,234 / 398 images respectively. From each split we discard instructions whose detailed question type is missing or appears in fewer than two samples (a stratification requirement), then draw a stratified subsample on detailed. The starting subsample sizes are 20,000 (train), 5,001 (val), and min(20000, available) = 12,574 (test); these targets were chosen to give the confidence probe a substantial training pool (20K samples is large enough for stable Optuna search across our five seeds while remaining tractable to extract hidden states for) and a validation pool roughly one quarter that size, while the test target is bounded by what GQA testdev makes available. After hash-based deduplication this yields **20,000 / 5,000 / 12,568** clean records spanning 104 / 101 / 91 detailed categories.

The standardized record stores the original GQA question verbatim, the short ground-truth answer, the scene image attached by image_id, and the detailed field as category. The five most frequent training categories are relS (10.1%), categoryRelS (9.3%), relO (8.9%), positionQuery (4.9%), and relChooser (4.0%). Question length is moderate (mean 8.8, median 8.0 tokens) and answers are essentially single-token (mean and median 1.0), reflecting the short-answer style that makes GQA suitable as a clean training distribution for confidence probes.

B.2.2 POPE (Object Hallucination Probing)

POPE (Li et al., 2023) probes object hallucination through yes/no questions of the form "Is there a {object} in the image?" grounded on COCO val2014 images (Lin et al., 2014), and serves as the near-ceiling binary detection task in VLCB. The lmms-lab/POPE test split contains 9,000 questions split equally among three negative-sampling regimes exposed as category: adversarial (objects co-occurring with scene content actually in the image, the hardest of the three), popular (frequent COCO objects), and random. The split is exactly answer-balanced (4,500 yes, 4,500 no). We retain all 9,000 questions; each becomes a one-token-answer record. The category distribution is reported in Table 7.

Table 7: POPE category distribution (test).

Category	Samples	%
adversarial	3,000	33.3
popular	3,000	33.3
random	3,000	33.3
Total	**9,000**	**100.0**

B.2.3 GMAI-MMBench (Medical Multimodal Reasoning)

GMAI-MMBench (Chen et al., 2024) is a large-scale multimodal medical benchmark covering radiology, pathology, dermatology, ophthalmology, and surgery, and is the principal high-stakes setting in VLCB. The HuggingFace train placeholder aggregates 25,831 questions but does not preserve the official split structure. The official splits are released as TSVs: GMAI_mm_bench_VAL.tsv (4,550 questions *with*

answers) and GMAI_mm_bench_TEST_part_*.tsv (∼21,281 questions, answers withheld for the leaderboard). Because the test partitions ship without ground truth, we use the official VAL split. One of the 4,550 raw images exceeds PIL's maximum allowed image size during decompression and is discarded, leaving **4,549** samples.

For each retained sample we (i) decode the base64 image, (ii) format the question by appending options "A) o_A B) o_B ..." (option E is omitted when null), and (iii) recover the option-letter ground truth by matching the textual category field against the option strings (case-insensitive, with a partial-match fallback). The standardized category field stores the source clinical VQA task (over 40 distinct values, with *Disease Diagnosis* dominating at 46.4% of the upstream pool). The answer-letter distribution after mapping is A:1,150 / B:1,175 / C:1,055 / D:977 / E:192, close to uniform across the four primary options with a long tail in option E.

B.2.4 MME-Finance (Financial Chart VQA)

MME-Finance (Gan et al., 2025) evaluates LVLMs on financial charts (candlestick, line, bar, table) rendered across four display styles (PC, photography, mobile vertical, mobile horizontal), and serves as the financial-document setting in VLCB. The release ships 1,171 question–image pairs across nine task categories. We retain only the six numerical, perceptual, OCR, and domain-knowledge categories that admit a single textual answer compatible with our pipeline, yielding **892** samples; the three categories dropped are open-ended explanation tasks for which a single textual ground truth is not well defined. The question field is taken verbatim, the answer preserves the dataset's free-form (often multi-line) reference text, and category stores the task_category field. The retained category distribution is reported in Table 8.

Table 8: MME-Finance task category distribution (test, after filtering).

Category	Samples	%
Spatial Awareness	229	25.7
OCR	178	20.0
Entity Recognition	163	18.3
Financial Knowledge	147	16.5
Accurate Numerical Calculation	133	14.9
Numerical Calculation	42	4.7
Total	**892**	**100.0**

B.2.5 MMMU-Pro (Multidisciplinary College-Level VQA)

MMMU-Pro (Yue et al., 2025) is a hardened version of MMMU spanning 30 college-level subjects across art, business, science, health, humanities, and engineering, with up to seven images per problem, and is the multi-choice reasoning setting in VLCB. We process the standard configurations released at 4 and 10 answer options (1,730 questions each); the screenshot-based vision configuration is also curated for completeness but is not used in the main evaluation suite. After hash-deduplication we retain **1,720** (4-option) and **1,725** (10-option) samples; the small drop relative to the 1,730 raw count reflects a handful of question–image pairs that hash identically across the two configurations. The category field is the joined string {subject}[SEP]{topic_difficulty}[SEP]{img_type joined by [LSEP]}, preserving subject (e.g. Math, Clinical Medicine), topic difficulty, and the list of contained image modalities. Subjects are approximately uniform (50–60 questions each); the full distribution is reported in Table 9. We include both 4-option and 10-option configurations because they share a question pool but differ in distractor count, which lets us examine whether confidence estimators degrade gracefully as the answer space widens.

B.2.6 LLaVA-in-the-Wild (Open-Ended Multimodal Instruction Following)

LLaVA-Bench-in-the-Wild (Liu et al., 2023b) is a 60-question benchmark of Open-ended visual dialogue spanning 24 in-the-wild images and three answer styles: conv (short conversational), detail (detailed description), and complex (multi-step reasoning). It is the smallest source in VLCB and serves as a stress test for confidence estimation under open-ended generation; per-bin reliability estimates on this dataset should be read with the small-sample caveat developed in §I.6. We pair each question with the GPT-4 reference response distributed alongside the benchmark and use it as the textual ground truth for our LLM-judge correctness grading (§B.5).

B.3 Final Train, Validation, and Test Assembly

The per-source artifacts are merged into three splits used downstream: a training set, a validation set, and a test set. The test split is the union of the seven test artifacts above, with duplicates by hash_id

Table 9: MMMU-Pro subject distribution (4-option and 10-option configurations are nearly identical pre-deduplication).

Subject	n	Subject	n	Subject	n
Agriculture	60	Diagnostics & Lab. Medicine	60	Public Health	58
Design	60	Psychology	60	Accounting	58
Finance	60	Clinical Medicine	59	Energy & Power	58
Physics	60	Biology	59	Pharmacy	57
Architecture & Engineering	60	Marketing	59	History	56
Electronics	60	Economics	59	Art Theory	55
Computer Science	60	Mechanical Engineering	59	Sociology	54
Math	60			Art	53
Music	60			Literature	52
Materials	60			Basic Medical Science	52
Chemistry	60			Geography	52
				Manage	50

removed (none were found, by design of the per-dataset hashing scheme). Training and validation are reserved exclusively from GQA. This is the operational realization of VLCB's design principle: a confidence estimator trained on a single, well-conditioned VQA distribution must generalize to test settings spanning medical imaging, financial chart understanding, multi-choice reasoning, and open-ended dialogue, none of which it has seen during training. Final sizes are reported in Table 10 and the test split's per-source composition in Table 11.

GQA and POPE together account for roughly 71% of the test split. This concentration reflects their established role as the two most widely used benchmarks in the LVLM confidence estimation and hallucination literature, where they have served as the de facto evaluation substrates for prior work; their prominence in VLCB therefore preserves comparability with that literature rather than reflecting a curation choice. The skew it introduces nonetheless motivates the two complementary aggregation modes used throughout our results: pooled aggregation (which weights every test sample equally and is therefore dominated by these two sources) and unweighted dataset averaging (which assigns equal weight to each of the seven sources irrespective of their sample counts). Both views are reported in Appendix I, and the contrast between them is what allows us to disentangle aggregate performance from per-domain robustness.

Quality control. Every assembled split is verified against a battery of assertions before any downstream use: every required field must be present and non-empty; every `hash_id` must be unique within its split; the pairwise intersection of `hash_ids` across train, val, and test must be empty; and the train and val splits must contain exclusively GQA samples. To further rule out any cross-source leakage that could in principle slip past a hash defined per-source, we additionally compute a *content-only* fingerprint over the canonicalized question text and a perceptual hash of the image, and intersect this fingerprint between (i) the training split and each of the seven test sources, and (ii) the validation split and each of the seven test sources. Across all 14 cross-source intersections, zero (image, question) pairs collide, confirming that no test sample has a content-equivalent twin in training or validation. As a looser diagnostic, we also intersect on image fingerprint alone: the only non-zero result is 5 GQA-train images that also appear in the POPE test set, attributable to the well-known Visual Genome–COCO image-pool overlap; the corresponding question texts differ (POPE's templated yes/no probes vs. GQA's scene-graph-derived compositional queries), so the (image, question) pairs do not collide. Image dimension statistics across source datasets are summarized in Table 12; the wide range of resolutions, from 43×26 pixels in MMMU-Pro to $6{,}824\times8{,}686$ pixels in GMAI-MMBench, reflects the heterogeneous visual content of the benchmark and motivates the 2,048-pixel downscaling policy applied at inference time (§B.4).

B.4 Response Generation

All five LVLMs are run on the full training, validation, and test splits under as uniform a generation protocol as their differing chat templates allow. Inference runs in `float32` with one exception: DeepSeek-VL2's vision encoder is numerically unstable outside `bfloat16` and is therefore loaded and evaluated in `bfloat16` throughout.

Table 10: VLCB split sizes and composition.

Split	Samples	Composition
VLCB_train_raw	20,000	GQA train (stratified on detailed)
VLCB_val_raw	5,000	GQA val (stratified on detailed)
VLCB_test_raw	30,514	7-source union, deduplicated by hash_id

Table 11: Per-source composition of VLCB_test_raw.

Source	Samples	%
GQA	12,568	41.19
POPE	9,000	29.50
GMAI-MMBench	4,549	14.91
MMMU-Pro (10-opt)	1,725	5.65
MMMU-Pro (4-opt)	1,720	5.64
MME-Finance	892	2.92
LLaVA-in-the-Wild	60	0.20
Total	**30,514**	**100.00**

This is a stability concession rather than a design choice and is the only point at which the generation conditions differ across models. Images whose larger dimension exceeds 2,048 pixels are downscaled while preserving aspect ratio; upsampling is never applied. Each sample is queried using the model's native chat and processor template, with generation fully deterministic (greedy decoding, max_new_tokens=64).

Prompt construction. Every model receives the same semantic instruction: answer the visual question briefly and completely, conditioned on the provided image. The instruction itself is held constant across models; only the delivery channel differs to accommodate each chat template. Qwen3-VL, Gemma-3, and InternVL3.5 use a dedicated system turn:

Qwen3-VL / Gemma-3 / InternVL3.5 — Prompt

System: You are a vision language assistant. Provide brief, complete answers.
User: {image} {question}

LLaVA-NeXT does not reliably honour a system turn, so the instruction is appended directly to the user message:

LLaVA-NeXT — Prompt

User: {image} {question}
Provide a brief, complete answer.

DeepSeek-VL2 uses a bespoke conversation schema where the system instruction is passed as an out-of-band processor argument rather than a chat turn:

DeepSeek-VL2 — Prompt

System (out-of-band): You are a vision language assistant. Provide brief, complete answers.
User: {image} {question}

B.5 Response Grading

We label every sample in VLCB, on every dataset, with a binary correctness score $y_i \in \{0, 1\}$ produced by a single LLM judge. The reason for using one protocol uniformly is that any alternative would mix grading rules across the benchmark: multiple-choice subsets could in principle be graded by string matching against the option key, and open-ended subsets necessarily require an LLM judge, but combining the two would mean systematic differences in measured confidence quality across datasets risk being attributable to grading artifacts rather than to genuine differences in model behavior. We therefore route the multiple-choice subsets (POPE, GMAI-MMBench, MMMU-Pro) through the judge as well, so the entire benchmark is graded under one rule. For the open-ended subsets (GQA short answers, MME-Finance, LLaVA-in-the-Wild) the LLM judge is in any case the only viable adjudicator, and this practice has been validated and widely adopted in recent calibration literature (Kapoor et al., 2024; Zhang et al., 2025; Chaudhry et al., 2024; Khanmohammadi et al., 2025, 2026).

Judge model and protocol. We use gpt-5-mini (with reasoning_effort=low) accessed through the OpenAI Responses API. For each sample, the judge receives the question, the ground-truth answer from the dataset, the LVLM's generated response, and the corresponding image as a separate multimodal input, and is asked to determine whether the generated answer is semantically equivalent to the ground truth. The reliability of using a powerful LLM for this constrained equivalence task has been demonstrated by Kapoor et al. (2024), who found that GPT-4 judgments exhibit a low

Table 12: Image dimension statistics per source dataset, computed over every (question, image) pair in the union of train, val, and test splits. Images are counted once per pair, so an image associated with multiple questions contributes its dimensions multiple times; this matches the distribution of image sizes the inference pipeline actually encounters during evaluation. Width and height are reported separately as images are non-square.

Dataset	N	W min	W max	W med	W mean	H min	H max	H med	H mean
GMAI-MMBench	4,549	55	6,824	512	857.4	54	8,686	434	681.0
GQA	37,568	72	1,280	500	523.3	51	1,280	406	433.5
LLaVA-Wild	60	505	4,800	1,214	1,416.9	374	3,203	1,152	1,173.4
MME-Finance	892	674	3,648	1,604	1,612.4	195	3,648	1,368	1,193.3
MMMU-Pro (10-opt)	1,725	43	2,954	602	725.3	26	2,560	357	484.0
MMMU-Pro (4-opt)	1,720	43	2,954	604	725.9	26	2,560	357	484.2
POPE	9,000	333	640	640	584.7	234	640	480	478.8

average absolute difference of 4.5% in accuracy estimation compared to human annotators. Building on their findings, and given the availability of more capable models since that study, we used `gpt-5-mini` to ensure high-quality labels at lower cost. The system and user prompts are given below.

Response Grading — System Prompt

You are an expert answer evaluator. Your task is to determine if a student's answer to a question is correct by comparing it to the ground truth answer.

1. Read the question carefully.
2. Compare the student's answer to the ground truth answer.
3. Consider semantic equivalence — answers that mean the same thing should be considered correct even if worded differently.
4. Return ONLY "yes" if the answer is correct, or "no" if it is incorrect.
5. Be lenient with minor variations in wording, capitalization, or punctuation.

Response Grading — User Prompt

Question: {question}
Ground Truth Answer: {ground_truth_answer}
Student Answer: {generated_response}
Is the student's answer correct? (yes/no):

Multimodal input format. The image is passed to the judge as a separate multimodal input alongside the textual user prompt rather than being serialized into the prompt text, which is why the template above contains no image placeholder.

The resulting correctness and response-length statistics for each LVLM, broken down by dataset and split, are reported in Table 13. Two patterns in the table are worth flagging because they shape how our results in §5 should be read. First, within-LVLM accuracy varies substantially across datasets: for Qwen3-VL-8B, for example, accuracy ranges from 37.2% on MMMU-Pro 10-option to 88.7% on POPE. This is the empirical reflection of the difficulty heterogeneity that motivates reporting an unweighted dataset average alongside pooled aggregation in Appendix I. Second, the cross-LVLM accuracy spread on the harder datasets, for example GMAI-MMBench (35.3% on LLaVA-NeXT to 60.0% on InternVL3.5), is the source of the variance against which the per-method confidence estimation results in the main paper must be read.

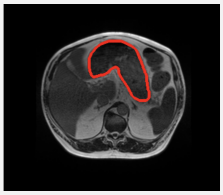
GQA — Compositional Visual Reasoning

Category: relChooser
Question: Is the brown horse to the right or to the left of the person that is standing on the road?
Answer: left

POPE — Object Hallucination Probing

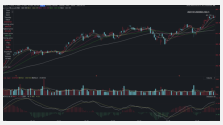
Category: adversarial
Question: Is there a spoon in the image?
Answer: no

GMAI-MMBench — Medical Multimodal Reasoning



Category: Organ Recognition – Abdomen
Question: This is an MRI image. Which option best describes the marked area?
A) heart, B) gallbladder, C) stomach, D) liver, E) necrotic tissue
Answer: C

MME-Finance — Financial Chart VQA



Category: Financial Knowledge
Question: What company is represented in the chart and what market is it listed on?
Answer: Microsoft Corporation, NASDAQ.

MMMU-Pro — Multidisciplinary College-Level VQA

Category: Agriculture[SEP]Hard[SEP]Photographs
Question: What could be the reason behind the browning on this potato leaf?
A) Overwatering, B) Mite feeding, C) Bacterial infection, D) Fungal infection,
E) Lack of sunlight, F) Don't know, G) Sunburn, H) Nutrient deficiency,
I) Ozone damage, J) Viral infection
Answer: B

LLaVA-in-the-Wild — Open-Ended Instruction Following

Category: conv
Question: Is there any strawberry-flavored yogurt in the fridge?
Answer: No. The fridge contains Fage non-fat yogurt, blueberry yogurt, and another yogurt cup with an unknown flavor.

Figure 3: One representative sample from each source dataset in VLCB. Each entry shows the image, task category, and question–answer pair.

Table 13: Per-VLM correctness breakdown by dataset and split. Train and validation splits contain only GQA, while the test split covers all seven source datasets. For each subset, we report total samples, correct and incorrect predictions (count and percentage), together with minimum, mean, and maximum word counts for the question and generated response. The aggregate row is reported only for the test split as **Total**.

Split	Dataset	Total	Correct (#, %)	Incorrect (#, %)	Q words	R words
			`Qwen/Qwen3-VL-8B-Instruct`			
Train	GQA	20,000	15,495 (77.5%)	4,505 (22.5%)	3/8.8/24	1/1.1/42
Val	GQA	5,000	3,785 (75.7%)	1,215 (24.3%)	3/8.8/23	1/1.1/36
Test	GMAI-MMBench	4,549	2,378 (52.3%)	2,171 (47.7%)	16/30.3/189	1/3.6/50
	GQA	12,568	8,670 (69.0%)	3,898 (31.0%)	3/8.5/25	1/1.1/53
	LLaVA-Wild	60	28 (46.7%)	32 (53.3%)	5/11.2/33	1/29.6/54
	MME-Finance	892	461 (51.7%)	431 (48.3%)	3/11.3/21	1/9.9/54
	MMMU-Pro (10-opt)	1,725	641 (37.2%)	1,084 (62.8%)	9/77.8/704	1/5.5/55
	MMMU-Pro (4-opt)	1,720	810 (47.1%)	910 (52.9%)	8/54.1/582	1/5.3/55
	POPE	9,000	7,979 (88.7%)	1,021 (11.3%)	7/7.2/8	1/1.0/1
	Total	**30,514**	**20,967 (68.7%)**	**9,547 (31.3%)**	**3/17.9/704**	**1/2.2/55**
			`llava-hf/llava-v1.6-vicuna-13b-hf`			
Train	GQA	20,000	15,850 (79.2%)	4,150 (20.8%)	3/8.8/24	1/2.7/56
Val	GQA	5,000	3,847 (76.9%)	1,153 (23.1%)	3/8.8/23	1/2.7/55
Test	GMAI-MMBench	4,549	1,605 (35.3%)	2,944 (64.7%)	16/30.3/189	2/9.2/53
	GQA	12,568	8,848 (70.4%)	3,720 (29.6%)	3/8.5/25	1/3.7/58
	LLaVA-Wild	60	18 (30.0%)	42 (70.0%)	5/11.2/33	1/36.5/56
	MME-Finance	892	194 (21.7%)	698 (78.3%)	3/11.3/21	1/21.3/58
	MMMU-Pro (10-opt)	1,725	318 (18.4%)	1,407 (81.6%)	9/77.8/704	1/22.9/60
	MMMU-Pro (4-opt)	1,720	470 (27.3%)	1,250 (72.7%)	8/54.1/582	1/22.5/60
	POPE	9,000	7,952 (88.4%)	1,048 (11.6%)	7/7.2/8	1/6.6/25
	Total	**30,514**	**19,405 (63.6%)**	**11,109 (36.4%)**	**3/17.9/704**	**1/8.1/60**
			`OpenGVLab/InternVL3_5-14B-HF`			
Train	GQA	20,000	15,155 (75.8%)	4,845 (24.2%)	3/8.8/24	1/29.6/62
Val	GQA	5,000	3,750 (75.0%)	1,250 (25.0%)	3/8.8/23	1/29.5/61
Test	GMAI-MMBench	4,549	2,730 (60.0%)	1,819 (40.0%)	16/30.3/189	1/10.4/57
	GQA	12,568	8,594 (68.4%)	3,974 (31.6%)	3/8.5/25	1/31.4/62
	LLaVA-Wild	60	15 (25.0%)	45 (75.0%)	5/11.2/33	8/42.1/58
	MME-Finance	892	460 (51.6%)	432 (48.4%)	3/11.3/21	1/25.2/59
	MMMU-Pro (10-opt)	1,724	541 (31.4%)	1,183 (68.6%)	9/77.9/704	1/33.5/59
	MMMU-Pro (4-opt)	1,720	654 (38.0%)	1,066 (62.0%)	8/54.1/582	1/35.4/60
	POPE	9,000	7,683 (85.4%)	1,317 (14.6%)	7/7.2/8	1/23.5/59
	Total	**30,513**	**20,677 (67.8%)**	**9,836 (32.2%)**	**3/17.9/704**	**1/26.1/62**
			`google/gemma-3-27b-it`			
Train	GQA	20,000	13,122 (65.6%)	6,878 (34.4%)	3/8.8/24	1/4.4/42
Val	GQA	5,000	3,244 (64.9%)	1,756 (35.1%)	3/8.8/23	1/4.4/40
Test	GMAI-MMBench	4,549	2,225 (48.9%)	2,324 (51.1%)	16/30.3/189	2/14.5/54
	GQA	12,568	7,479 (59.5%)	5,089 (40.5%)	3/8.5/25	1/5.3/52
	LLaVA-Wild	60	31 (51.7%)	29 (48.3%)	5/11.2/33	1/31.0/56
	MME-Finance	892	381 (42.7%)	511 (57.3%)	3/11.3/21	1/11.8/51
	MMMU-Pro (10-opt)	1,717	618 (36.0%)	1,099 (64.0%)	9/76.3/558	1/21.9/55
	MMMU-Pro (4-opt)	1,718	822 (47.8%)	896 (52.2%)	8/53.6/508	1/20.4/55
	POPE	9,000	7,565 (84.1%)	1,435 (15.9%)	7/7.2/8	1/1.3/23
	Total	**30,504**	**19,121 (62.7%)**	**11,383 (37.3%)**	**3/17.8/558**	**1/7.5/56**
			`deepseek-ai/deepseek-vl2`			
Train	GQA	20,000	12,883 (64.4%)	7,117 (35.6%)	3/8.8/24	1/8.4/64
Val	GQA	5,000	3,154 (63.1%)	1,846 (36.9%)	3/8.8/23	1/8.3/64
Test	GMAI-MMBench	4,549	1,670 (36.7%)	2,879 (63.3%)	16/30.3/189	0/11.2/61
	GQA	12,568	6,757 (53.8%)	5,811 (46.2%)	3/8.5/25	1/9.1/64

(Continued on next page)

(Continued from previous page)

Split	Dataset	Total	Correct (#, %)	Incorrect (#, %)	Q words	R words
	LLaVA-Wild	60	16 (26.7%)	44 (73.3%)	5/11.2/33	1/32.9/64
	MME-Finance	892	265 (29.7%)	627 (70.3%)	3/11.3/21	0/7.3/64
	MMMU-Pro (10-opt)	1,725	185 (10.7%)	1,540 (89.3%)	9/77.8/704	0/31.8/64
	MMMU-Pro (4-opt)	1,720	316 (18.4%)	1,404 (81.6%)	8/54.1/582	0/30.5/64
	POPE	9,000	7,573 (84.1%)	1,427 (15.9%)	7/7.2/8	1/6.5/35
	Total	**30,514**	**16,782 (55.0%)**	**13,732 (45.0%)**	**3/17.9/704**	**0/11.1/64**

B.6 Benchmark Availability and Licensing

VLCB is a composite resource over seven public datasets, each governed by its own license; licenses range from permissive (Apache 2.0, MIT, CC BY) to restrictive (CC BY-NC-SA, research-use-only). We release the components needed to reconstruct it: the per-source curation code and the aggregator that assembles the three splits, the generation driver and the LLM-judge grading code, and the deterministic hash_id construction. Any user who has independently obtained the source datasets from their official distributors can therefore reproduce the VLCB splits bit-for-bit by running our code; the assembled benchmark is not redistributed as a single archive because doing so would conflict with the more restrictive source licenses. All released code is licensed under MIT. The reconstructed VLCB benchmark is a derivative work and inherits the most restrictive terms of its constituent sources; it is therefore intended for **non-commercial research use only** and is subject to all applicable ShareAlike provisions inherited from GMAI-MMBench. Users are solely responsible for acquiring the source datasets from their official distributors and for adhering to their original license terms.

C Baseline Confidence Estimation Methods

This appendix documents the seven established confidence estimation baselines benchmarked in this work. Each method operates on the language model backbone and reads one of three signals to estimate confidence in a generated answer: output logits, verbalized confidence scores, or hidden states. We organize the methods into three families along this axis: prompt-based methods (§C.1), internal-state probing methods (§C.2), and internal-stability methods (§C.3). For trainable methods, we adhere to the architectural descriptions of the original publications wherever they are specified, with deviations flagged in the relevant subsections. All baselines operate on the generation pass described in §B.4; hidden states and logits are extracted from the same frozen LVLM checkpoints listed in Tables 4–5, with no additional fine-tuning of the underlying LVLM. Throughout this appendix, each method is referred to by the canonical name we use in the rest of the paper; Table 14 also lists the short-form abbreviation used in figure legends and per-method labels in the results.

C.1 Prompt-Based Methods

Prompt-based methods treat the LVLM as an externally queried system, deriving confidence through prompt design rather than access to internal representations. The three methods in this family differ in how they extract a confidence signal: P(True) reads a token logit from a self-evaluation query, Self-Probing parses a verbalized confidence score, and Prompt Ensembles aggregates output probabilities across paraphrased questions. All three are training-free, and all three require at least one additional inference pass beyond the original answer-generation pass (Self-Probing and P(True) require one additional generation each, Prompt Ensembles requires ten).

C.1.1 P(True)

P(True), introduced by Kadavath et al. (2022), assesses the LVLM's self-evaluation of its own generated answer through a token-logit readout on a binary self-query. P(True) does not access hidden representations or activations; it does require token-level output scores for the auxiliary response token, which is a standard decoding output exposed by most local inference frameworks and by several API providers via log-probability endpoints, so no internal model access is implied. After the LVLM produces a response to the original visual question, the (image, question, generated answer) triple is fed back with the following uncertainty query:

P(True) — Uncertainty Query

Is the proposed answer correct?
A) no
B) yes
Reply with A or B only.
Answer:

Implementation. The model generates a single token under greedy decoding and the output logits at that position are recorded. To be robust to tokenizer variation across LVLMs, we collect logits for all A/B token variants (upper- and lowercase, with and without a leading space, with and without surrounding parentheses) and retain the maximum within each equivalence class. Let ℓ_A and ℓ_B denote the resulting scalar logits for the "no" and "yes" options respectively. The final confidence score is

$$p_{\text{True}} = \text{softmax}([\ell_A, \ell_B])[1]. \quad (6)$$

Method	Citation	Model Access	Input Signal	Training Required
		Prompt-Based		
P(True)	(Kadavath et al., 2022)	Token-logit access	Output token scores	No
Self-Probing	(Xiong et al., 2024)	Black-box	Verbalized confidence	No
Prompt Ensemble (PE)	(Zollo and Zemel, 2025)	Black-box	Aggregated output probabilities	No
		Internal-State Probing		
P(I Know)	(Kadavath et al., 2022)	White-box (hidden states)	First-token final-layer hidden state	Yes
SAPLMA	(Azaria and Mitchell, 2023)	White-box (hidden states)	Last-token final-layer hidden state	Yes
InternalInspector (I^2)	(Beigi et al., 2024)	White-box (all-layer hidden states)	Per-layer activation, attention, and FFN states	Yes
		Internal-Stability		
CCPS	(Khanmohammadi et al., 2025)	White-box (hidden states + gradients)	Perturbation trajectory statistics	Yes

Table 14: Overview of the seven confidence estimation baselines benchmarked in this work, organized by the family of signal they exploit, the degree of model access they require, and whether they involve any training.

C.1.2 Self-Probing (SP)

Self-Probing, proposed by Xiong et al. (2024), prompts the LVLM to verbalize a numerical confidence in its own answer through a free-text generation pass. After the LVLM produces a response, it is queried with:

Self-Probing — Verbalization Prompt

Question: {question}
Possible Answer: {generated_answer}

Q: How likely is the above answer to be correct? give your confidence in the following format:
Confidence: <number from 0 to 100>%
Note: The confidence indicates how likely you think the answer is true.

Implementation. The verbalized response is parsed with a regular expression to extract the confidence value, which is then normalized to $[0, 1]$.

C.1.3 Prompt Ensembles (PE)

Prompt Ensembles, formalized by Zollo and Zemel (2025), estimates confidence by averaging predictions across N meaning-preserving rewrites of the input question. The LVLM independently answers the original question and each of the N paraphrases under greedy decoding, yielding $N + 1$ generated responses per sample. A per-prompt confidence is computed as the length-normalized sequence likelihood of the response, and the ensemble confidence is the arithmetic mean of the $N + 1$ per-prompt scores. The intuition is that an answer the model is sure of should remain stable across semantically equivalent rephrasings, while an answer driven by surface-level cues should fluctuate; averaging across paraphrases therefore acts as a soft consistency check at the output-distribution level.

PE — System Prompt

You are an expert linguist and domain specialist generating alternative phrasings for a Visual Question Answering (VQA) task. Your goal is to generate variations of a question while preserving its exact semantic meaning and expected answer.

Instructions:

- Generate the requested number of alternative ways to phrase the question.
- **Preserve the Answer:** Ensure that for any given image, the answer to your new questions would be *identical* to the answer of the original question.
- **Maintain Domain Precision:**
  - If the question is **Medical**, preserve the correct clinical terminology.
  - If the question is **Financial**, keep the technical intent clear.
- **Maintain Question Type:**
  - If the original is a **Yes/No** question, your rephrasings must remain Yes/No questions.
  - If the original asks for a **Count**, your rephrasings must ask for a number.
- **Multiple Choice Handling (Crucial):**
  - If the original question includes multiple options (e.g., A, B, C, D):
    1. Rephrase **only** the question stem (the text asking the question).
    2. **Append the EXACT same options** in the **EXACT same order** to the end of your rephrased question.
    3. Do **not** shuffle, reword, or modify the options in any way.
- **Allowed Changes:** You may vary word order, sentence structure, and use strict synonyms for the question text.
- **Prohibited Changes:** Do **not** add new constraints, remove location details, or introduce ambiguity.

Output Format:

Each rephrased question should be wrapped in numbered tags like this:

```
[question_1] Rephrased question stem? (A)
Option 1 (B) Option 2... [/question_1]
[question_2] Rephrased question stem? (A)
Option 1 (B) Option 2... [/question_2]
...and so on for each question.
```

PE — User Prompt

Original Question: `'{question}'`
Please generate {N} alternative phrasings for this question.
Output Format:

```
[question_1] ... [/question_1]
[question_2] ... [/question_2]
...
[question_{N}] ... [/question_{N}]
```

Implementation. We use $N = 10$, the smallest ensemble size at which Zollo and Zemel (2025) report near-saturated calibration gains, balancing the per-paraphrase inference cost against the marginal benefit of adding another rewrite. Paraphrases are generated via the OpenAI `gpt-5-mini` API with reasoning effort set to `medium`, under the system and user prompts shown above. For each of the eleven (image, question) pairs (the original plus ten paraphrases) we record per-token log-probabilities $\{\log p_t\}_{t=1}^{T}$ of the generated response and compute a per-prompt confidence as the length-normalized sequence likelihood,

$$c^{(i)} = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\log p_t\right), \tag{7}$$

equivalent to the geometric mean of per-token probabilities. This follows standard practice for sequence-likelihood confidence in LLM uncertainty estimation (Kuhn et al., 2023; Malinin and Gales, 2021). The ensemble confidence is the arithmetic mean of the eleven per-prompt scores, $c^{\text{PE}} = \frac{1}{N+1}\sum_{i=0}^{N} c^{(i)}$.

C.2 Internal-State Probing Methods

Probing methods access the LVLM's internal activations and train a lightweight classifier head to predict answer correctness from static snapshots of those activations. The three methods in this family differ in which forward pass they read from (prompt-only or extended through the generated response) and in how the resulting hidden states are aggregated; the specifics are given per method below. Across all three methods the LVLM weights are held fixed; with methods using early stopping on the composite validation score defined in Appendix E. All reported metrics are computed on the test set.

C.2.1 P(I Know) (P(IK))

P(I Know), also from Kadavath et al. (2022), estimates the probability that the LVLM will produce a correct answer *before* any response is generated. We implement P(IK) as a lightweight classifier head on the frozen LVLM: during the extraction forward pass we capture the final-layer hidden state at the last token of the input context (the system prompt followed by the user message containing the image and the question), yielding a vector $h \in \mathbb{R}^d$ that summarizes the model's representation of the full input at the moment it would otherwise begin generating. The probe therefore reads only the prompt, with no generated response in the forward pass; this contrasts with SAPLMA and InternalInspector, which both extend the pass through the generated response and read from its final token. P(IK) is the natural architectural counterpart to `BICR`: both train an MLP over a single hidden-state vector from a prompt-only pass at the same token position, and `BICR`'s improvement over P(IK) in the main results (§5) is therefore attributable to the blank-image ranking signal rather than to any architectural difference.

Implementation. The classifier is a multi-layer perceptron with ReLU activations, trained with `BCEWithLogitsLoss` using a positive-class reweighting factor to correct for class imbalance and optimized with Adam. Because the original publication does not specify head hyperparameters at the LVLM scale we work with, we tune them per LVLM via an Optuna search on the validation set; the full search space and training protocol are documented in Appendix F. At inference, the confidence score is $\sigma(f_\theta(h))$.

C.2.2 SAPLMA

SAPLMA (Azaria and Mitchell, 2023) trains a lightweight feedforward classifier on the final-layer hidden state at the last token of the model's generated response to predict answer correctness. The original method reads this hidden state at the end of the generated text alone, which is appropriate in text-only settings where the response is a self-contained statement: "the capital of the US is DC" carries its own truth conditions and a representation of just that statement is enough to assess it. LVLM responses are typically not self-contained in this way. A one-word answer like "red" to "what color is this" is uninterpretable in isolation; whether it is correct depends entirely on the image and the question that prompted it. We therefore extract the hidden state at the last token of the full sequence consisting of the input context (system prompt, im-

age, question) followed by the generated response, so that the probe reads a representation reflecting the joint (image, question, response) context after the model has committed to its answer. This is the only departure from the original SAPLMA convention.

Implementation. The classifier is a four-layer MLP with hidden widths $d \to 256 \to 128 \to 64 \to 1$, ReLU activations, and a final linear projection to a scalar logit, following the architecture of Azaria and Mitchell (2023). Training uses `BCEWithLogitsLoss` with a positive-class reweighting factor to correct for class imbalance, optimized with Adam at batch size 32 for at most 200 epochs with early stopping on the composite validation score (Appendix E) with patience 20. To reduce seed sensitivity we train five independent classifiers with seeds $\{23, 42, 137, 2024, 3407\}$ and report the mean of all metrics across runs. At inference, the confidence score is $\sigma(f_\theta(h))$.

C.2.3 InternalInspector (I^2)

Beigi et al. (2024) argue that useful correctness signal is distributed across the full depth of the model rather than concentrated in a single layer. For each sample we extract three per-layer representations at the last token of the full sequence consisting of the input context (system prompt, image, question) followed by the generated response, via forward hooks on every transformer block: the post-residual activation state $h^{(l)}$, the pre-residual multi-head self-attention output $a^{(l)}$, and the pre-residual feed-forward output $m^{(l)}$. Stacking these across L layers yields a per-sample tensor of shape $[L, d, 3]$, which is treated as a 3-channel image with L rows (one per layer), d columns (one per hidden dimension), and one channel for each of the three state types. We benchmark the strongest variant reported in the original work, a CNN-based encoder over all three state types, which outperforms alternatives that use a Transformer encoder or a subset of state types. The trainable parameter count of this variant is fixed at approximately 11.3M regardless of the underlying LVLM, because the CNN encoder reduces the input to a fixed spatial footprint via adaptive pooling before the projection head; the full parameter accounting is in Appendix G.

Implementation. The $[L, d, 3]$ tensor is passed through a ResNet18 CNN encoder (stem followed by four residual stages with channel progression $64 \to 128 \to 256 \to 512$) and adaptive average pooling, yielding a 512-dimensional embedding. A linear projection $512 \to 128$ followed by a four-layer MLP classifier ($128 \to 256 \to 128 \to 64 \to 1$) with ReLU activations and dropout 0.1 produces the correctness logit. The encoder and classifier are trained jointly from scratch using a supervised contrastive loss combined with `BCEWithLogitsLoss` (with positive-class reweighting for class imbalance) following Beigi et al. (2024, Eq. 4), with Adam at learning rate 10^{-3}, weight decay 10^{-4}, and contrastive temperature $\tau = 0.1$. Training runs for at most 200 epochs with early stopping on the composite validation score (Appendix E) with patience 20.

C.3 Internal-Stability Methods

Stability-based methods probe the LVLM's internal representations not by reading static activations but by measuring how those representations *respond* to controlled perturbations. The confidence signal is representational robustness: hidden states that retain their predictive content under targeted intervention are treated as reliable, while those whose predictions shift easily under the same intervention are treated as fragile. We evaluate one such method.

C.3.1 CCPS

CCPS (Khanmohammadi et al., 2025) estimates confidence by probing the stability of the LVLM's final hidden states under targeted adversarial perturbations: hidden states behind correct predictions should resist small interventions, while those behind incorrect ones should shift easily. For each token in the generated response, the hidden state is perturbed along the unit-normalized gradient direction in S=5 equal-sized steps up to $\varepsilon_{\max}$=20.0 (matching the original publication's settings), and per-token statistics are recorded across three groups, namely original-state features, perturbation-trajectory features, and comparison features quantifying the distributional shift between original and perturbed states (Table 15), yielding a 75-channel per-token feature sequence over the full response.

Implementation. CCPS uses a two-stage head. *Stage 1* pre-trains a convolutional encoder, `Conv1d`$(75 \to 64, k{=}3) \to$ ReLU $\to$ `Conv1d`$(64 \to 32, k{=}3) \to$ ReLU $\to$ `AdaptiveMaxPool1d` $\to$ `Linear`$(32 \to 16)$, with a margin-based contrastive loss (margin 1.0,

class-agnostic by construction) for 5,000 steps. *Stage 2* appends $\texttt{Linear}(16 \to 32) \to \text{ReLU} \to \texttt{Linear}(32 \to 2)$ and jointly fine-tunes under cross-entropy (with positive-class reweighting for class imbalance) for a further 5,000 steps. Both stages use Adam at learning rate 10^{-4}, weight decay 0.1, and batch size 32. At inference, the confidence score is the softmax probability that the generated response is correct (i.e., the probability mass on the positive class of the binary classifier).

D Evaluation Metrics

We evaluate confidence estimation quality along two complementary axes: *calibration*, which measures how well predicted confidence scores reflect true correctness frequencies, and *discrimination*, which measures how well they separate correct from incorrect predictions. The metrics reported throughout this work are organized below.

D.1 Calibration Metrics

D.1.1 Expected Calibration Error (ECE)

A well-calibrated confidence estimator should assign a score of p to predictions that are correct p of the time in expectation. We measure calibration via ECE, which partitions all n samples into b equal-width bins $\{B_j\}_{j=1}^{b}$ over $[0, 1]$ and computes the weighted average absolute deviation between mean predicted confidence and empirical accuracy:

$$\text{ECE} = \sum_{j=1}^{b} \frac{|B_j|}{n} \left|\text{conf}(B_j) - \text{acc}(B_j)\right|$$

where $\text{conf}(B_j)$ and $\text{acc}(B_j)$ denote the average confidence and observed accuracy within bin B_j, respectively. We use $b = 10$ equal-width bins throughout. Lower ECE indicates better alignment between predicted scores and actual correctness rates.

D.1.2 Brier Score (BS)

The Brier Score measures the mean squared error between each predicted confidence p_k and the binary correctness label $o_k \in \{0, 1\}$:

$$\text{BS} = \frac{1}{N} \sum_{k=1}^{N} (p_k - o_k)^2$$

It reflects both calibration and the ability to assign informative probabilities, penalizing estimators that are overconfident, underconfident, or stuck near 0.5 regardless of correctness. Lower values reflect higher overall reliability.

D.2 Discrimination Metrics

The first two discrimination metrics (Accuracy and F1) are threshold-based and require binarizing the confidence score. We use a fixed default threshold of 0.5 on the probe's sigmoid output for every method, with no per-method or per-dataset threshold tuning. This avoids any test-set-derived threshold selection and keeps the comparison protocol uniform across methods. The remaining two metrics (AUCPR and AUROC) summarize performance across all thresholds and require no such selection. Model selection during training is governed by a separate composite validation score described in Appendix E and is not used to tune any test-time threshold.

D.2.1 Accuracy (ACC)

Accuracy measures the proportion of samples for which the binarized confidence prediction agrees with the ground-truth correctness label:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Here, the positive class is a correct LVLM response and the negative class is an incorrect one. We report Accuracy to contextualize the difficulty of each evaluation setting and to enable direct comparison with methods that report threshold-based performance.

D.2.2 F1 Score (F1)

F1 is the harmonic mean of precision ($\text{TP}/(\text{TP} + \text{FP})$) and recall ($\text{TP}/(\text{TP} + \text{FN})$), again with the positive class defined as a correct LVLM response:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Relative to Accuracy, F1 is more informative when the correct/incorrect class distribution is skewed, since it weights the joint quality of identifying correct predictions and avoiding false alarms on incorrect ones, while ignoring true negatives entirely.

D.2.3 Area Under the Precision–Recall Curve (AUCPR)

AUCPR summarizes precision against recall across all confidence thresholds, with the positive class again defined as a correct LVLM response. Relative to AUROC, it places greater weight on performance over the positive class, making it more informative when the class distribution is skewed.

D.2.4 Area Under the ROC Curve (AUROC)

AUROC measures the probability that a randomly drawn correct prediction receives a higher confidence score than a randomly drawn incorrect one, $P(s^+ > s^-)$, where s^+ and s^- denote scores assigned to correct and incorrect predictions respectively. It is threshold-independent and relatively insensitive to class prevalence compared with threshold-based metrics. A score of 1.0 reflects perfect separation; 0.5 corresponds to chance.

E Validation Monitoring and Model Selection

Training the methods in our benchmark raises a non-trivial model-selection question, since confidence estimation is a two-objective problem: a confidence score must *discriminate* correct from incorrect answers and must also be *calibrated* in an absolute sense. A training run that minimizes a cross-entropy loss does not directly favour either property and can drift between them from epoch to epoch. To make checkpoint selection principled and consistent across methods, we monitor a single composite validation score throughout training and use it as the unifying signal for model selection across every trainable method in our benchmark. All five trainable methods (SAPLMA, P(I Know), InternalInspector, CCPS, and our proposed `BICR`) use this score for early stopping with a patience of 20 validation steps applied uniformly across all five. Two of these (P(I Know) and `BICR`) additionally use the same score as the Optuna optimization objective; see Appendix F. Across all five methods, training uses positive-class reweighted binary cross-entropy in which the positive label corresponds to a correct LVLM response. Class imbalance is handled by setting $w_+ = n_-/n_+$, where n_+ and n_- are the counts of correct and incorrect samples in the training split respectively, so that the loss contribution from each class is balanced regardless of which class is the majority.

Following the protocol established by Khanmohammadi et al. (2026), the composite validation score is a convex combination of AUROC and $(1 - \text{ECE})$:

$$\text{CompositeScore} = \alpha \cdot \text{AUROC} + (1 - \alpha) \cdot (1 - \text{ECE}) \quad (8)$$

and we fix $\alpha = 0.6$ across every trainable method in the benchmark. We adopt this weighting for the same reason articulated by the original work: ranking correctness reliably is the primary practical requirement of a confidence estimator in deployment, so a slight preference for AUROC over calibration error is warranted, but a score that is discriminatively strong yet grossly miscalibrated cannot be interpreted as a probability and is therefore of limited utility. The $0.6/0.4$ split penalizes miscalibration sharply enough to discourage degenerate solutions that collapse onto a narrow confidence distribution while still allowing ranking quality to break ties between otherwise comparable checkpoints.

Use in training. For all five trainable methods, we evaluate the composite score on the validation split at every validation step and retain the checkpoint with the highest composite score as the final model. The early-stopping patience of 20 validation steps means training terminates when the composite has not improved for 20 consecutive checks. For the two Optuna-tuned methods (P(I Know) and `BICR`), the composite score at the best epoch of each trial is returned as the trial's objective value, so both per-trial checkpoint selection and cross-trial hyperparameter selection are driven by the same quantity. `BICR`

F Hyperparameter Search with Optuna

Of the seven baselines documented in Appendix C, three are inference-only and therefore not subject to training (P(True), Self-Probing, Prompt Ensembles), and three further methods (SAPLMA, InternalInspector, CCPS) are trained with the exact architectures and hyperparameters prescribed by their original publications. Our proposed method `BICR` and one baseline, P(I Know), are trained with an Optuna hyperparameter search. This appendix describes the search protocol shared by both methods and the search space each one explores.

F.1 Optimization Protocol

We use the `Optuna` framework (Akiba et al., 2019) with a Tree-structured Parzen Estimator (TPE) sampler, seeded by the same random seed that drives the training data pipeline so that search behaviour is reproducible. The protocol below applies identically to `BICR` and P(I Know).

Each (method, LVLM, seed) tuple is optimized for 50 trials, with five independent seeds $\{23, 42, 137, 2024, 3407\}$ run per (method, LVLM) pair, and all downstream evaluation metrics reported as the mean across seeds. Each trial trains for at most 200 epochs at batch size 32, with

early stopping on the composite validation score (Appendix E) at patience 20. The trial's objective value is the composite score at its best epoch, and the trial with the highest objective is selected as the final configuration for that (LVLM, seed). To accelerate the search, Optuna's Median Pruner is applied with 5 start-up trials, 10 warm-up steps, and an interval of 5 steps, terminating unpromising trials early on the same intermediate composite score. To prevent model capacity from being conflated with raw parameter count when comparing architectural efficiency, all configurations are constrained to a maximum of 5,000,000 trainable parameters; any trial suggesting a model outside this budget is pruned before training.

F.2 Hyperparameter Search Space

The search spaces for BICR and P(I Know) are summarized in Table 16. Both methods tune the MLP classifier architecture (depth, width, dropout) and optimizer settings (learning rate, weight decay). BICR additionally tunes three loss-coefficient hyperparameters that control the auxiliary training objectives coupling its two extraction views ($\mathbf{h}_{\text{base}}$, $\mathbf{h}_{\text{blank}}$): β (weight on the Brier calibration term $\mathcal{L}_{\text{brier}}$ on $\mathbf{h}_{\text{base}}$), λ (weight on the visual-grounding ranking loss $\mathcal{L}_{\text{rank}}$ that contrasts $\mathbf{h}_{\text{base}}$ against $\mathbf{h}_{\text{blank}}$), and γ (the ranking-loss margin in probability space; see Eq. 3 in §4). The fixed-architecture methods (SAPLMA, InternalInspector, CCPS) are not listed in the table, since their architectures and hyperparameters are taken verbatim from the original publications and are documented in Appendix C.

Note on the loss-coefficient search. The four shared rows in Table 16 cover the classifier architecture (depth, width, dropout) and optimizer settings, and are identical between BICR and P(I Know) so that the two methods compete on the same architectural and optimization footing. BICR alone tunes the three loss-coefficient hyperparameters in the second row group (β, λ, γ), since these control the auxiliary objectives that couple its two extraction views and have no analogue in P(I Know)'s BCE-only training. The analysis of the Optuna-selected values for these three coefficients across LVLMs and seeds is provided in Appendix H.3.

G Analysis of Additional Trainable Parameters

This appendix quantifies and compares the *additional* learnable parameters introduced by each evaluated confidence estimation method, including our proposed BICR, when applied to a frozen base LVLM. We first report the architectural dimensions of the LVLMs that drive the parameter counts of the linear-probe methods (§G.1), then provide the formulas for the additional trainable parameters of each method (§G.2), and finally report the exact parameter counts used in our experiments (§G.3), followed by a short discussion (§G.4). All counts are of *trainable* parameters (i.e., parameters returned by `nn.Module.parameters()` with `requires_grad=True`); batch-normalization running statistics and other registered buffers are excluded. All counts include biases unless otherwise noted.

G.1 Base LVLM Architectural Parameters

The key architectural dimensions of the base LVLMs used in this study that influence the number of trainable parameters of probe-style methods are summarized in Table 17: the language-model hidden size d_h and the number of decoder layers L. We also list the total number of parameters of each LVLM (read directly from its `model.safetensors.index.json`); the percentages reported in §G.3 use this column as denominator. The full model cards (vision-encoder details, number of attention heads, etc.) are given in Appendix A.

G.2 Formulation of Additional Trainable Parameters

Table 18 lists the trainable components and the closed-form parameter expressions for each method. For InternalInspector and CCPS the count is independent of the base LVLM, so no formula in d_h or L is needed. For the Optuna-tuned methods (P(I Know) and BICR), the selected depth and widths of the classifier head differ per (LVLM, seed); we write $(H_1, \ldots, H_k)$ for the tuple of hidden widths that Optuna selects for a given run.

Note on InternalInspector. The ResNet18 encoder used by InternalInspector takes a 3-channel input corresponding to the stacked (activation, attention, feed-forward) states at each layer. The spatial dimensions of its input are (L, d_h), but because every downstream layer is followed by adap-

tive average pooling to a fixed $(1, 1)$ spatial footprint before the $512 \rightarrow 128$ projection, the total trainable-parameter count is independent of both L and d_h.

Note on CCPS. CCPS uses a two-stage training pipeline: a contrastive encoder is first pre-trained with 21,168 parameters (Stage 1), and then a 610-parameter classification head (Linear(16→32) + Linear(32→2)) is attached on top of the encoder, with the entire stack fine-tuned end-to-end (Stage 2). The final deployed model is the Stage 2 classifier checkpoint, which contains all encoder weights plus the classification head, totalling 21,778 parameters. The head operates on a per-token sequence of 75 trajectory features computed from the frozen LVLM's hidden states, and no part of the dimension scales with d_h; the parameter count is therefore identical across all LVLMs.

Note on BICR. BICR uses two views: the base hidden state $\mathbf{h}_{\text{base}}$ and the blank-image hidden state $\mathbf{h}_{\text{blank}}$. Both views pass through the *same* MLP at training time, but only $\mathbf{h}_{\text{base}}$ is processed at inference. The blank view serves exclusively as a training-time regularizer through the ranking loss $\mathcal{L}_{\text{rank}}$, adding zero parameters and zero inference cost beyond the standard MLP probe. The parameter formula is therefore identical to that of P(I Know); the difference between the two methods lies entirely in the training objective, not the architecture.

G.3 Exact Additional Trainable Parameter Counts

Table 19 reports the exact number of additional trainable parameters introduced by each method when applied to each base LVLM, per benchmark seed. Each cell reports the trainable parameter count of the saved checkpoint and, in parentheses, the count expressed as a percentage of the total parameter count of the corresponding base LVLM (Table 17).

The fixed-architecture methods (SAPLMA, InternalInspector, CCPS) have a single count per (method, LVLM); we still tabulate all five seeds so that readers can verify the per-seed determinism of the reconstruction, and so that the layout of the table is uniform across methods. The two Optuna-tuned methods (P(I Know) and BICR) re-run the architecture search independently for each (method, LVLM, seed) tuple, as documented in Appendix F, and consequently the selected `classifier_layers` (and therefore the trainable-parameter count) can differ across seeds; for these two methods the cell additionally reports the Optuna-selected `classifier_layers` tuple in light text. The bottom MEAN row of each method's block reports the mean over the five seeds, of both the parameter count and the percentage of the base-LVLM's parameters.

G.4 Discussion

Table 19 makes the capacity of each benchmarked confidence estimator transparent and directly comparable. Three qualitative regimes are apparent. First, the three prompt-based methods (P(True), Self-Probing, Prompt Ensembles) introduce no trainable parameters at all and rely entirely on the frozen LVLM's own output behaviour. Second, CCPS occupies an extreme position at the other end of the spectrum of trained methods: its fixed two-stage convolutional head totals only 21,778 parameters regardless of the base LVLM (well under 0.001% of every backbone), because its input is already a compact 75-channel trajectory-feature sequence rather than the high-dimensional LVLM hidden state. Third, the remaining trainable methods (SAPLMA, InternalInspector, P(I Know), and BICR) occupy a moderate regime of a few hundred thousand to roughly eleven million trainable parameters, with the exact count driven by the LVLM hidden size d_h (SAPLMA, P(I Know), BICR) or by the fixed convolutional encoder alone (InternalInspector).

Within the moderate regime, InternalInspector is consistently the heaviest at 11,316,417 parameters on every LVLM, ranging from 0.04% of Gemma-3-27B and DeepSeek-VL2 up to 0.13% of Qwen3-VL-8B, because its ResNet18-style CNN encoder dominates the count. BICR is substantially lighter: averaging the parameter counts across the five seeds, BICR's MLP head ranges from 642,561 parameters on Qwen3-VL-8B to 2,753,537 on InternVL3.5-14B, which translates to a $4.1\times$ to $17.6\times$ reduction relative to InternalInspector across the five LVLMs (the smallest gap appearing on InternVL3.5-14B and the largest on Qwen3-VL-8B). Compared to its architectural counterpart P(I Know), which uses an identical MLP formula but tends toward larger Optuna-selected widths, BICR is lighter on four of the five LVLMs (notably $4.7\times$ lighter on Qwen3-VL-8B, where Optuna selects $(128, 64)$ for BICR on four of five seeds while selecting $(1024, 512)$ or larger for P(I Know)

on the highest-parameter seeds), and is slightly heavier than P(I Know) only on InternVL3.5-14B, where Optuna selects $(512, 256)$ for BICR on every seed while P(I Know) tends toward narrower (512) heads. SAPLMA is the only method that is consistently lighter than BICR, at 0.7M–1.4M parameters depending on d_h, owing to its fixed shallow architecture. The pattern across BICR's selections is consistent with its design: the training-time ranking loss $\mathcal{L}_{\text{rank}}$ provides a stronger learning signal than BCE alone, allowing smaller architectures to reach competitive performance, so the capacity advantage comes from the training objective rather than from the model size.

A separate point worth flagging is that BICR's blank view contributes zero parameters to the deployed model, since only $\mathbf{h}_{\text{base}}$ is processed at inference; the blank-view hidden state is consumed exclusively at training time by $\mathcal{L}_{\text{rank}}$. BICR is therefore a strictly parameter-equivalent alternative to single-view probes like P(I Know) at deployment, with the performance improvement coming entirely from the training signal.

Taken as a whole, every trainable method in our benchmark adds at most $\approx 0.13\%$ of the smallest base LVLM's parameter count, supporting our framing of these confidence estimators as genuinely lightweight additions to a frozen LVLM rather than as meaningful contributions to the deployed model's size or inference cost.

H Design Validation and Ablation Analysis

This appendix presents the empirical evidence underlying each design choice in BICR. All experiments follow the benchmark evaluation protocol documented in Appendix F: 5 LVLMs $\times$ 5 seeds $\times$ 50 Optuna trials per (LVLM, seed) tuple, with metrics averaged across seeds (per-LVLM) or across LVLMs (cross-LVLM). Statistical significance is assessed via paired Wilcoxon signed-rank tests over the 25 (LVLM, seed) observations.

H.1 Loss Component Ablation

BICR's training objective (Eq. 4) consists of three terms: $\mathcal{L}_{\text{bce}}$, $\mathcal{L}_{\text{brier}}$, and $\mathcal{L}_{\text{rank}}$. We ablate each by removing it while keeping all other components unchanged, yielding four configurations: the full BICR model, $-\mathcal{L}_{\text{brier}}$ (removing the calibration term), $-\mathcal{L}_{\text{rank}}$ (removing the visual grounding ranking; note that this also removes the need for $\mathbf{h}_{\text{blank}}$ entirely, as the blank view is only used by $\mathcal{L}_{\text{rank}}$), and $\mathcal{L}_{\text{bce}}$-only (removing both auxiliary terms).

Cross-LVLM results. Table 20 reports the cross-LVLM average metrics for each ablation variant.

Key findings. $\mathcal{L}_{\text{rank}}$ is the critical component: removing it degrades AUROC by 3.3 points ($p < 0.001$, paired Wilcoxon) and AUCPR by 1.9 points ($p < 0.001$). This confirms that the blank-image contrastive signal is the primary driver of BICR's discriminative advantage. $\mathcal{L}_{\text{brier}}$ contributes a more modest calibration benefit on the cross-LVLM average (ΔECE $= -1.4$, $p = 0.059$; ΔBS $= -0.6$, $p = 0.019$). Removing both auxiliary losses ($\mathcal{L}_{\text{bce}}$-only) produces the worst configuration on every discrimination metric, with the degradation highly significant ($p < 0.005$).

Statistical significance, pooled aggregation. Table 21 reports the significance test results for each ablation comparison under cross-LVLM pooled aggregation.

Statistical significance, equal-weight aggregation. Table 22 reports the same Wilcoxon tests under per-dataset equal-weight aggregation. Two findings track the cross-LVLM pooled ablation: removing $\mathcal{L}_{\text{rank}}$ produces highly significant degradation on every metric, and removing both auxiliary losses ($\mathcal{L}_{\text{bce}}$-only) produces significant degradation on calibration metrics. Under this aggregation, removing $\mathcal{L}_{\text{brier}}$ alone is not statistically significant on any metric (all $p > 0.05$), suggesting that the Brier term's contribution is concentrated on the larger datasets that dominate the pooled aggregation rather than on the smaller datasets that get equal weight here. Removing both auxiliary losses also weakens significance on AUCPR (n.s.) and AUROC ($p = 0.045$) compared to the much stronger pooled-aggregation effect, again indicating that auxiliary-loss benefits concentrate on the dominant datasets.

Cluster-aware significance. The paired Wilcoxon tests above treat each (LVLM, seed) tuple as an independent observation, but the 5 seeds within an LVLM share the same frozen weights and the same test set, so the truly independent unit is the LVLM. To verify the loss-ablation findings under a cluster-aware protocol, Table 23 reports a cluster bootstrap (10,000 resamples) over LVLM-level seed-means

with Holm-Bonferroni correction across the 4 metrics. The results sharpen rather than overturn the n=25 conclusions: $\mathcal{L}_{\text{rank}}$'s discriminative contribution remains highly significant on AUCPR, AUROC, and BS ($p < 0.001$ on each), and the $\mathcal{L}_{\text{bce}}$-only comparison shows the same pattern with comparable effect sizes. The contribution of $\mathcal{L}_{\text{brier}}$ alone is the only place the picture softens: its ablation does not reach Holm-corrected significance under cluster-aware testing, consistent with the n=25 unweighted-aggregation result that $\mathcal{L}_{\text{brier}}$'s benefit concentrates on specific datasets rather than as a uniform across-LVLM effect. ECE comparisons are also non-significant under this conservative protocol because the across-LVLM ECE distribution is heavy-tailed (Table 24), but the corresponding mean deltas are all in the expected direction.

Per-LVLM breakdown. Table 24 presents the loss ablation results for each LVLM individually. The discrimination story is uniform: $\mathcal{L}_{\text{rank}}$ improves AUROC and AUCPR on every LVLM, with the largest gains on LLaVA-NeXT (ΔAUROC $= +6.0$) and DeepSeek-VL2 (ΔAUROC $= +4.1$). Calibration trade-offs are more nuanced: on Qwen3-VL-8B, removing $\mathcal{L}_{\text{rank}}$ yields a lower ECE than the full model (4.9 vs. 8.9), and on Gemma-3-27B removing $\mathcal{L}_{\text{brier}}$ yields a slightly lower ECE (6.4 vs. 7.0). These per-LVLM ECE patterns reflect the fact that the auxiliary terms target the cross-LVLM average, where the calibration gain is consistent (Table 20); on individual LVLMs, the rank loss occasionally trades a small amount of calibration error for its substantial discrimination gain.

H.2 What $\mathcal{L}_{\text{rank}}$ Teaches the Probe

Beyond improving aggregate metrics, $\mathcal{L}_{\text{rank}}$ fundamentally changes the probe's learned behavior. We analyze this by comparing the per-sample confidence outputs of the full BICR model against the $\mathcal{L}_{\text{bce}}$-only baseline, using all 25 runs (5 LVLMs $\times$ 5 seeds).

Confidence separation. Table 25 compares the mean confidence assigned to correct versus incorrect samples.

Without $\mathcal{L}_{\text{rank}}$, incorrect samples receive substantially higher confidence (49.2 vs. 41.6), reducing the separation by 20%. $\mathcal{L}_{\text{rank}}$ selectively suppresses confidence for incorrect predictions while maintaining high confidence for correct ones.

Calibration. The effect is most pronounced in the high-confidence range. Table 26 compares the reliability diagram bins where overconfidence is most harmful.

The Full vs $-\mathcal{L}_{\text{rank}}$ comparison isolates the ranking loss's contribution at each confidence level (with $\mathcal{L}_{\text{brier}}$ present in both). The high-confidence range $[0.6, 0.9)$ is where the rank loss matters most: removing $\mathcal{L}_{\text{rank}}$ produces severe overconfidence (a 75% prediction corresponds to 64% empirical accuracy, gap $= 10.7\%$), while keeping it in place tracks the diagonal much more closely (75% prediction corresponds to 72% accuracy, gap $= 3.1\%$). Across the four high-confidence bins $[0.6, 1.0]$ the gap reduction ranges from $2.7\times$ (in $[0.9, 1.0]$) to $5.2\times$ (in $[0.6, 0.7)$), with the largest improvements concentrated in precisely the confidence range where overconfidence is most consequential for downstream decision-making. The $\mathcal{L}_{\text{bce}}$-only column confirms that removing both auxiliary losses does not produce calibration meaningfully worse than removing $\mathcal{L}_{\text{rank}}$ alone in this range, indicating that $\mathcal{L}_{\text{rank}}$ is the dominant driver of high-confidence calibration in BICR. On two of the low-confidence bins ($[0.0, 0.1)$ and $[0.4, 0.5)$), $-\mathcal{L}_{\text{rank}}$ lands within 1–3 gap points of Full and occasionally edges it; this reflects the asymmetric design intent of $\mathcal{L}_{\text{rank}}$, which targets overconfidence at high scores and is therefore expected to dominate in the high-confidence range where overconfidence is the practical risk, rather than in the low-confidence range where the failure mode is the opposite (underconfidence: predicted scores below empirical accuracy in all variants).

Figure 4 visualizes this effect across the full reliability curve. Each panel shows one reliability curve per seed (5 curves total, drawn translucently), and a grey histogram in the lower portion of each panel showing the distribution of predicted confidence values pooled across all 5 LVLMs and 5 seeds. Moving from left to right (Full $\rightarrow -$Brier $\rightarrow -$Rank $\rightarrow$ BCE-only), the curves progressively bow further away from the diagonal in the low-to-mid confidence range, where empirical accuracy sits above predicted confidence (i.e., the probe is underconfident in that range, predicting low scores for samples that turn out correct more often than the score implies). The full BICR model tracks the diagonal closely across the full range; without $\mathcal{L}_{\text{rank}}$ the curves develop a clear bow upward in the $[0, 0.5]$ region. The pooled histograms tell a complementary story: the full model spreads prob-

ability mass across the full range while the ablated variants (especially BCE-only) concentrate mass toward the high-confidence end.

Score utilization. Without $\mathcal{L}_{\text{rank}}$, the probe concentrates scores in the high-confidence range: 66.8% of all samples receive scores above 0.5, and only 1.9% fall below 0.1. With $\mathcal{L}_{\text{rank}}$, the distribution broadens (60.1% above 0.5, 4.1% below 0.1), indicating that the probe has learned to use the full probability range and express genuine uncertainty when the model's prediction is unreliable. This broader utilization is a hallmark of well-calibrated confidence estimators.

H.3 Optimized Hyperparameter Analysis

The Optuna search jointly optimizes the loss weights β ($\mathcal{L}_{\text{brier}}$), λ ($\mathcal{L}_{\text{rank}}$), and the margin γ alongside the architectural and optimizer hyperparameters. Table 27 reports the mean and range of these values across the 5 seeds for each LVLM, providing insight into the stability and LVLM-dependence of the optimized configurations.

Three patterns are noteworthy. First, the Brier weight β is consistently selected in a moderate range (mean 0.16–0.35), confirming that Optuna finds the calibration term beneficial but secondary to the classification objective. Second, the rank weight λ is always positive and non-trivial (mean 0.07–0.20), demonstrating that the optimizer consistently allocates weight to the ranking signal across all LVLMs. Although one seed on LLaVA-NeXT-13B does select $\lambda = 0.01$, no LVLM's seed-level mean approaches the lower bound, supporting that $\mathcal{L}_{\text{rank}}$ provides genuine training signal rather than being an artifact of the search space. Third, the margin γ trends toward larger values for the LVLMs with the lowest baseline correctness rates (DeepSeek-VL2 at $\gamma = 0.17$, baseline 55.0% correct; Gemma-3-27B at $\gamma = 0.17$, baseline 62.7% correct), and toward smaller values for the LVLM with the highest correctness rate (Qwen3-VL-8B at $\gamma = 0.09$, baseline 68.7% correct), although the relationship is not perfectly monotonic across the middle of the range. The Optuna-selected architectures vary across LVLMs from compact $(128, 64)$ heads on Qwen3-VL-8B to wider $(512, 256)$ on InternVL3.5-14B and DeepSeek-VL2, without a clean correspondence to LVLM hidden size; the architecture choice appears to depend more on interactions with the loss-coefficient settings than on d_h alone.

H.4 Choice of Null Image

BICR uses a solid black image as the null visual input $v_\varnothing$ throughout the main experiments. This choice is load-bearing: the rank loss $\mathcal{L}_{\text{rank}}$ pushes $\sigma(W^\top \mathbf{h}_{\text{base}})$ above $\sigma(W^\top \mathbf{h}_{\text{blank}})$ for correctly answered samples, so the visual content of the null directly shapes the supervisory signal. To test whether the result is specific to black or whether BICR is robust to any visually-impoverished null, we systematically compare five null-image strategies, each chosen to isolate a specific axis along which a null can differ from the base image: **black** (current default; minimum visual signal, dark uniform field), **white** (minimum visual signal, bright uniform field; tests whether the effect comes from luminance or uniformity), **Gaussian noise** (uniform-random pixels; tests whether BICR responds to the absence of information or the absence of image-like structure), **blurred original** (low-frequency content preserved via Gaussian blur with radius 50, high-frequency content stripped; tests whether BICR responds to loss of detail or to total image absence), and **pixel-shuffled original** (per-pixel permutation; preserves the color histogram but destroys all spatial layout, testing whether the signal comes from spatial structure or color statistics).

Experimental setup. For each of the four new null types we re-extract the BICR feature representation from scratch and re-train the BICR probe over the same five seeds $\{23, 42, 137, 2024, 3407\}$ used in the main experiments, then compare to the existing black-baseline checkpoints on the held-out test split. The two stochastic generators (Gaussian noise and pixel-shuffled) use a deterministic per-sample seed $\sigma = \text{hash}_{32}(h_{\text{id}}) \oplus s$, where h_{id} is the sample's hash identifier and $s = 42$ is a global null seed, so the same sample produces the same null pixels across reruns. To bound compute cost, this experiment is restricted to a single backbone (`Qwen3-VL-8B-Instruct`). All other forward-pass settings, training hyperparameters, Optuna budget (50 trials per seed), and evaluation protocol match the main BICR pipeline.

Test-set performance per null type. Table 29 reports paired Wilcoxon signed-rank tests comparing each new null type against the black baseline, paired by seed ($n = 5$). With only five paired observations, the smallest p-value the Wilcoxon test can produce is 0.0625, attained when all five seeds

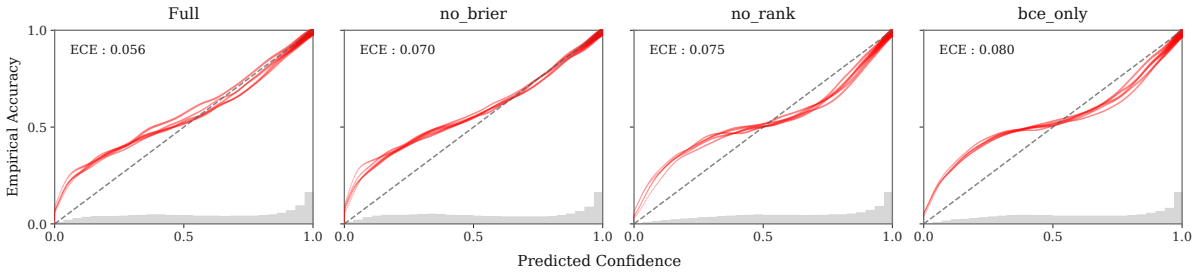


Figure 4: Per-seed reliability diagrams for each ablation variant. Each panel shows one reliability curve per seed (5 translucent red curves) and a grey histogram in the lower portion of the panel showing the distribution of predicted confidence values pooled across all 5 LVLMs and 5 seeds (25 runs combined). The dashed diagonal marks perfect calibration. Cross-LVLM ECE values are shown in each panel: Full BICR achieves the lowest ECE (0.056); removing $\mathcal{L}_{\text{rank}}$ raises ECE to 0.075, and removing both auxiliary losses raises it further to 0.080.

agree on the direction of the effect. Cells at this floor (marked †) therefore represent the strongest possible evidence at this sample size: black wins (or loses) on every single seed. Cells above the floor (Gaussian noise on ECE at $p = 0.125$, blurred on ECE at $p = 0.188$) indicate that fewer than all five seeds agreed on the direction.

Black is the best of the five strategies on every metric. The black baseline achieves the best mean performance on all six metrics across all four alternative null types. The margin is large in absolute terms: AUROC drops from 0.801 (black) to between 0.678 and 0.722 for the four alternatives, a 7.9 to 12.3 point gap; AUCPR collapses from 0.901 to roughly 0.51–0.54; F1 drops from 0.787 to roughly 0.33–0.39. Calibration also worsens: ECE roughly doubles ($0.089 \rightarrow 0.12$–0.13) and Brier rises by ≈ 0.04. The direction of the effect is unanimous across the five paired seeds for every metric in $\{\text{Brier}, \text{Acc}, \text{F1}, \text{AUCPR}, \text{AUROC}\}$ and for ECE under the white and pixel-shuffled comparisons (Wilcoxon $p = 0.0625$, the floor for $n = 5$). In the two cells where the ECE comparison falls slightly above the floor (Gaussian noise at $p = 0.125$ and blurred at $p = 0.188$), the mean still favors black.

Why the alternatives fail, by axis of variation. The four alternative null types form an interpretable failure pattern that maps cleanly onto the four axes the comparison was designed to probe. **Luminance vs. uniformity:** white and black are both uniform fills, yet white loses 12.3 AUROC points relative to black. Uniformity alone is therefore not what makes black useful as a null; the LVLM produces meaningfully different hidden states for white than for black, and the latter are more useful as a contrast point. **Information content vs. image-likeness:** replacing the image with high-entropy random pixels (Gaussian noise) does not help, dropping AUROC by 9.7 points relative to black. The blank view should be *information-poor* rather than *information-different*; adding visual entropy without spatial structure is closer to "a different image" than to "no image." **High-frequency detail vs. total absence:** heavy blurring strips edges, text, and object structure but preserves low-frequency content (broad color fields, average illumination), and still costs 10.0 AUROC points relative to black, suggesting that the residual low-frequency content is enough for the LVLM to maintain a non-null hidden state. **Spatial structure vs. color statistics:** pixel-shuffling preserves the image's full color histogram while scrambling all spatial layout. This is the least bad of the four alternatives (only 7.9 AUROC points worse than black), consistent with spatial structure carrying the bulk of the image-likeness signal; once layout is destroyed, the LVLM behaves closer to the all-black case, although the gap to black remains significant on every metric.

Design implication. The simple solid-black null both produces the most distinct $\mathbf{h}_{\text{blank}}$ from $\mathbf{h}_{\text{base}}$ and provides the strongest training signal for the rank loss. The ablation supports the design choice in BICR: the null view should be information-poor in an absolute sense, not merely information-different. A null that still carries any image-like content, whether high-entropy noise, low-pass-filtered original, or shuffled color statistics, gives the LVLM enough to produce a non-null representation that the probe cannot exploit as effectively as it can the all-black contrast.

I Extended Results

This appendix provides the complete set of results underlying the main-text summary in §5. All trained methods report the mean across five seeds $\{23, 42, 137, 2024, 3407\}$, each with 50 Optuna hyperparameter trials. Metrics are computed on the shared subset of test samples present in all methods for a given LVLM, obtained by intersecting test `hash_ids` across every confidence estimation method evaluated on that LVLM; per-LVLM and per-dataset shared counts are reported in Table 30. The intersection accounts for the small number of samples that drop from individual methods for benign reasons: prompt-based methods occasionally produce a malformed parse on a sample whose generated text does not match the expected response format, and Self-Probing requires a second LVLM forward pass that can fail on individual samples due to context-length overflow. Restricting evaluation to the intersection enforces apples-to-apples comparison: every metric, significance test, and figure in this appendix and in the main paper is computed on the same per-LVLM sample set across all methods. Best values per metric are **bolded**. Reported metrics are: Expected Calibration Error (ECE), Brier Score (BS), Accuracy (ACC), F1 Score (F1), Area Under Precision–Recall Curve (AUCPR), and Area Under ROC Curve (AUROC).

I.1 Per-LVLM Pooled Performance

Table 31 reports the pooled aggregate performance of each method on each LVLM. In pooled evaluation, all test samples are combined into a single set and metrics are computed over the full shared subset per LVLM. `BICR` achieves the best AUCPR and AUROC on every LVLM (five of five), and the best calibration (ECE, BS) on three of five (LLaVA-NeXT-13B, InternVL3.5-14B, and DeepSeek-VL2); on the remaining two LVLMs the best ECE belongs to InternalInspector (Qwen3-VL-8B and Gemma-3-27B).

Original State Features	
original_log_prob_actual	Log-probability of the actual token under the model's unperturbed output distribution.
original_prob_actual	Probability of the actual token under the unperturbed distribution.
original_logit_actual	Raw logit of the actual token prior to any perturbation.
original_prob_argmax	Highest probability assigned to any token by the unperturbed model.
original_logit_argmax	Highest logit value assigned to any token prior to perturbation.
original_entropy	Entropy of the unperturbed predictive distribution: $-\sum_i P_{\text{orig}}(i) \log P_{\text{orig}}(i)$.
original_margin_logit_top1_top2	Logit gap between the top-1 and top-2 tokens before perturbation.
original_margin_prob_top1_top2	Probability gap between the top-1 and top-2 tokens before perturbation.
original_norm_logits_L2	L2 norm of the unperturbed logit vector.
original_std_logits	Standard deviation of the unperturbed logit values.
original_norm_hidden_state_L2	L2 norm of the unperturbed last hidden state vector.
is_actual_token_original_argmax	Binary indicator of whether the actual token is the argmax under the unperturbed model.
Perturbation Trajectory Features	
jacobian_norm_token	L2 norm of the Jacobian of the token's log-probability with respect to the hidden state, measuring local sensitivity.
epsilon_to_flip_token	Smallest perturbation magnitude along the gradient direction sufficient to change the argmax prediction.
pei_value_token	Perturbation Energy Integral (PEI): cumulative normalized drop in the actual token's log-probability across all perturbation steps.
Comparison Features (Original vs. Perturbed)	
perturbed_log_prob_actual	Log-probability of the actual token after hidden-state perturbation.
perturbed_prob_actual	Probability of the actual token after perturbation.
perturbed_logit_actual	Logit of the actual token after perturbation.
perturbed_prob_argmax	Highest probability assigned to any token after perturbation.
perturbed_logit_argmax	Highest logit value after perturbation.
perturbed_entropy	Entropy of the perturbed predictive distribution.
perturbed_margin_logit_top1_top2	Logit gap between top-1 and top-2 tokens after perturbation.
perturbed_norm_logits_L2	L2 norm of the perturbed logit vector.
delta_log_prob_actual_from_original	Absolute drop in log-probability of the actual token after perturbation.
did_argmax_change_from_original	Binary indicator of whether the argmax token shifted after perturbation.
kl_div_perturbed_from_original	KL divergence from the original to the perturbed output distribution.
js_div_perturbed_from_original	Jensen-Shannon divergence between the original and perturbed distributions.
cosine_sim_logits_perturbed_to_original	Cosine similarity between logit vectors before and after perturbation.
cosine_sim_hidden_perturbed_to_original	Cosine similarity between hidden-state vectors before and after perturbation.
l2_dist_hidden_perturbed_from_original	L2 distance between hidden-state vectors before and after perturbation.

Table 15: Feature groups and definitions used by CCPS (Khanmohammadi et al., 2025) to characterize hidden-state stability under targeted perturbation.

Table 16: Hyperparameter search space for the Optuna-tuned methods. Both methods share the classifier-architecture and optimizer search space; `BICR` additionally tunes its three loss coefficients. Notation `a,b` inside a set denotes a single categorical choice corresponding to an MLP with hidden widths (a, b).

Component	**Search space**
Shared (BICR and P(I Know))	
`classifier_layers`	{`0`; `256`; `512`; `128,64`; `256,128`; `512,256`; `1024,512`; `1024,512,256`}
`classifier_dropout`	$\{0.0,\ 0.1,\ 0.3,\ 0.5\}$
`learning_rate`	$[10^{-5},\ 10^{-3}]$, log-uniform
`weight_decay`	$[10^{-6},\ 10^{-3}]$, log-uniform
BICR-only (ours)	
β (Brier weight)	$[0.0,\ 0.5]$, uniform
λ (rank weight)	$[0.01,\ 0.3]$, uniform
γ (margin)	$[0.05,\ 0.25]$, uniform

Table 17: Architectural dimensions and total parameter counts of the base LVLMs used in this study. The total-parameters column is obtained by summing the product of tensor shape dimensions over every tensor listed in `model.safetensors.index.json` and is used as the denominator for the % column in Table 19.

Base LVLM	d_h	L	**Total parameters**
`Qwen/Qwen3-VL-8B-Instruct`	4,096	36	8,767,123,696 (8.77 B)
`llava-hf/llava-v1.6-vicuna-13b-hf`	5,120	40	13,351,499,776 (13.35 B)
`OpenGVLab/InternVL3_5-14B-HF`	5,120	40	15,119,523,840 (15.12 B)
`google/gemma-3-27b-it`	5,376	62	27,432,406,640 (27.43 B)
`deepseek-ai/deepseek-vl2`	2,560	30	27,480,134,248 (27.48 B)

Table 18: Formulas for additional trainable parameters introduced by each method. d_h is the LVLM hidden size from Table 17; $(H_1, \ldots, H_k)$ is the Optuna-selected tuple of hidden widths of the classifier head for a given run.

Method	**Trainable component(s)**	**Formula (incl. biases)**
P(True)	None (prompting only)	0
Self-Probing	None (prompting only)	0
Prompt Ensembles	None (inference only)	0
SAPLMA	MLP $d_h \to 256 \to 128 \to 64 \to 1$	$256\, d_h + 41{,}473$
InternalInspector	ResNet18-style CNN encoder (3-channel input) + projection $512 \to 128$ + MLP $128 \to 256 \to 128 \to 64 \to 1$	11,316,417 (independent of d_h, L)
CCPS	**Stage 1:** Conv1d($75{\to}64, k{=}3$) + Conv1d($64{\to}32, k{=}3$) + Linear($32{\to}16$) **Stage 2:** encoder fine-tuned jointly + Linear($16{\to}32$) + Linear($32{\to}2$)	21,778 (independent of d_h, L; encoder embedded in classifier)
P(I Know)	MLP $d_h \to H_1 \to \cdots \to H_k \to 1$	$d_h H_1 + H_1 + \sum_{i=1}^{k-1} \left(H_i H_{i+1} + H_{i+1}\right) + (H_k + 1)$
`BICR` (ours)	MLP $d_h \to H_1 \to \cdots \to H_k \to 1$ (shared across base and blank views; blank view used only at training time via $\mathcal{L}_{\text{rank}}$)	$d_h H_1 + H_1 + \sum_{i=1}^{k-1} \left(H_i H_{i+1} + H_{i+1}\right) + (H_k + 1)$

Table 19: Per-seed trainable-parameter counts and percentage of total base-LVLM parameters for every confidence-estimation method in our benchmark. Each cell reports the trainable parameter count of the saved checkpoint and, in parentheses, the count as a percentage of the total parameter count of the corresponding base LVLM. Percentages are rounded to two decimals; cells displayed as 0.00% are non-zero but below 0.005%. SAPLMA, InternalInspector, and CCPS use fixed architectures, so their counts are identical across the five benchmark seeds $\{23, 42, 137, 2024, 3407\}$; for the Optuna-tuned methods (P(I Know) and BICR), counts vary across seeds and the cell additionally reports the Optuna-selected `classifier_layers` tuple. The bottom MEAN row of each method's block is the across-seed mean of both the parameter count and the percentage.

Method	**Seed (VLM total)**	**Qwen3-VL-8B** *8.77 B*	**LLaVA-Next-13B** *13.35 B*	**InternVL3.5-14B** *15.12 B*	**Gemma-3-27B** *27.43 B*	**DeepSeek-VL2** *27.48 B*
Prompt-only methods (no trainable parameters)						
P(True)	—	0	0	0	0	0
Self-Probing	—	0	0	0	0	0
Prompt Ensemble	—	0	0	0	0	0
SAPLMA	23	1,090,049 (0.01%)	1,352,193 (0.01%)	1,352,193 (0.01%)	1,417,729 (0.01%)	696,833 (0.00%)
	42	1,090,049 (0.01%)	1,352,193 (0.01%)	1,352,193 (0.01%)	1,417,729 (0.01%)	696,833 (0.00%)
	137	1,090,049 (0.01%)	1,352,193 (0.01%)	1,352,193 (0.01%)	1,417,729 (0.01%)	696,833 (0.00%)
	2024	1,090,049 (0.01%)	1,352,193 (0.01%)	1,352,193 (0.01%)	1,417,729 (0.01%)	696,833 (0.00%)
	3407	1,090,049 (0.01%)	1,352,193 (0.01%)	1,352,193 (0.01%)	1,417,729 (0.01%)	696,833 (0.00%)
	MEAN	**1,090,049 (0.01%)**	**1,352,193 (0.01%)**	**1,352,193 (0.01%)**	**1,417,729 (0.01%)**	**696,833 (0.00%)**
InternalInspector	23	11,316,417 (0.13%)	11,316,417 (0.08%)	11,316,417 (0.07%)	11,316,417 (0.04%)	11,316,417 (0.04%)
	42	11,316,417 (0.13%)	11,316,417 (0.08%)	11,316,417 (0.07%)	11,316,417 (0.04%)	11,316,417 (0.04%)
	137	11,316,417 (0.13%)	11,316,417 (0.08%)	11,316,417 (0.07%)	11,316,417 (0.04%)	11,316,417 (0.04%)
	2024	11,316,417 (0.13%)	11,316,417 (0.08%)	11,316,417 (0.07%)	11,316,417 (0.04%)	11,316,417 (0.04%)
	3407	11,316,417 (0.13%)	11,316,417 (0.08%)	11,316,417 (0.07%)	11,316,417 (0.04%)	11,316,417 (0.04%)
	MEAN	**11,316,417 (0.13%)**	**11,316,417 (0.08%)**	**11,316,417 (0.07%)**	**11,316,417 (0.04%)**	**11,316,417 (0.04%)**
CCPS	23	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)
	42	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)
	137	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)
	2024	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)
	3407	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)	21,778 (0.00%)
	MEAN	**21,778 (0.00%)**	**21,778 (0.00%)**	**21,778 (0.00%)**	**21,778 (0.00%)**	**21,778 (0.00%)**
P(I Know)	23	4,720,641 (0.05%; 1024,512)	2,622,465 (0.02%; 512)	2,622,465 (0.02%; 512)	2,753,537 (0.01%; 512)	1,311,745 (0.00%; 512)
	42	1,081,857 (0.01%; 256,128)	2,622,465 (0.02%; 512)	2,622,465 (0.02%; 512)	1,376,769 (0.01%; 256)	1,311,745 (0.00%; 512)
	137	2,229,249 (0.03%; 512,256)	2,753,537 (0.02%; 512,256)	1,311,233 (0.01%; 256)	1,409,537 (0.01%; 256,128)	1,442,817 (0.01%; 512,256)
	2024	2,098,177 (0.02%; 512)	1,311,233 (0.01%; 256)	2,753,537 (0.02%; 512,256)	2,753,537 (0.01%; 512)	1,442,817 (0.01%; 512,256)
	3407	4,851,713 (0.06%; 1024,512,256)	2,622,465 (0.02%; 512)	2,622,465 (0.02%; 512)	2,753,537 (0.01%; 512)	3,278,849 (0.01%; 1024,512,256)
	MEAN	**2,996,327 (0.03%)**	**2,386,433 (0.02%)**	**2,386,433 (0.02%)**	**2,209,383 (0.01%)**	**1,757,595 (0.01%)**
BICR (ours)	23	532,737 (0.01%; 128,64)	2,753,537 (0.02%; 512,256)	2,753,537 (0.02%; 512,256)	1,409,537 (0.01%; 256,128)	1,442,817 (0.01%; 512,256)
	42	1,081,857 (0.01%; 256,128)	1,344,001 (0.01%; 256,128)	2,753,537 (0.02%; 512,256)	1,409,537 (0.01%; 256,128)	1,442,817 (0.01%; 512,256)
	137	532,737 (0.01%; 128,64)	1,344,001 (0.01%; 256,128)	2,753,537 (0.02%; 512,256)	696,577 (0.00%; 128,64)	1,442,817 (0.01%; 512,256)
	2024	532,737 (0.01%; 128,64)	663,809 (0.00%; 128,64)	2,753,537 (0.02%; 512,256)	1,409,537 (0.01%; 256,128)	1,442,817 (0.01%; 512,256)
	3407	532,737 (0.01%; 128,64)	663,809 (0.00%; 128,64)	2,753,537 (0.02%; 512,256)	2,884,609 (0.01%; 512,256)	1,442,817 (0.01%; 512,256)
	MEAN	**642,561 (0.01%)**	**1,353,831 (0.01%)**	**2,753,537 (0.02%)**	**1,561,959 (0.01%)**	**1,442,817 (0.01%)**

Table 20: Loss ablation for BICR, averaged across 5 LVLMs $\times$ 5 seeds. Each row removes one or more loss components. ΔAUROC is the change relative to the full model. Best values in **bold**.

Variant	**ECE↓**	**BS↓**	**AUCPR↑**	**AUROC↑**	**Δ**
Full (BICR)	**7.1**	**18.4**	**87.4**	**78.6**	—
$-\mathcal{L}_{\text{brier}}$	8.5	19.0	87.1	78.0	−0.6
$-\mathcal{L}_{\text{rank}}$	8.1	19.6	85.5	75.3	−3.3
$\mathcal{L}_{\text{bce}}$ only	9.2	19.9	85.5	75.2	−3.4

Table 21: Statistical significance of loss ablation (paired Wilcoxon, n=25). Each cell shows the p-value; ***, **, * denote $p < 0.001$, $p < 0.01$, $p < 0.05$; "n.s." denotes $p \geq 0.05$.

Comparison	**ECE**	**BS**	**AUCPR**	**AUROC**
Full vs. $-\mathcal{L}_{\text{brier}}$	n.s.	0.019^{*}	n.s.	n.s.
Full vs. $-\mathcal{L}_{\text{rank}}$	n.s.	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
Full vs. $\mathcal{L}_{\text{bce}}$ only	0.004^{**}	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$

Table 22: Statistical significance of loss ablation under per-dataset equal-weight aggregation (paired Wilcoxon, n=25). Same conventions as Table 21.

Comparison	**ECE**	**BS**	**AUCPR**	**AUROC**
Full vs. $-\mathcal{L}_{\text{brier}}$	n.s.	n.s.	n.s.	n.s.
Full vs. $-\mathcal{L}_{\text{rank}}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{**}	$<0.001^{***}$
Full vs. $\mathcal{L}_{\text{bce}}$ only	$<0.001^{***}$	$<0.001^{***}$	n.s.	0.045^{*}

Table 23: Cluster-aware significance of BICR's loss ablation under pooled aggregation. Each row reports the mean per-LVLM delta (Full BICR minus ablated variant) across 5 LVLMs, with significance assessed by a cluster bootstrap (10,000 resamples) over LVLM-level seed-means and Holm-Bonferroni correction across the 4 metrics within each row. Significance level conventions match Table 39.

Comparison	**ECE ↓**		**BS ↓**		**AUCPR ↑**		**AUROC ↑**	
	Mean Δ	p	**Mean Δ**	p	**Mean Δ**	p	**Mean Δ**	p
Full vs $-\mathcal{L}_{\text{brier}}$	-0.0140	n.s.	-0.0061	n.s.	+0.0032	n.s.	+0.0059	n.s.
Full vs $-\mathcal{L}_{\text{rank}}$	-0.0104	n.s.	-0.0120	$<0.001^{***}$	+0.0189	$<0.001^{***}$	+0.0328	$<0.001^{***}$
Full vs $\mathcal{L}_{\text{bce}}$ only	-0.0205	n.s.	-0.0147	$<0.001^{***}$	+0.0193	$<0.001^{***}$	+0.0334	$<0.001^{***}$

Table 24: Per-LVLM loss ablation results (mean across 5 seeds). Best value per LVLM per metric in **bold**.

LVLM	**Variant**	**ECE↓**	**BS↓**	**AUCPR↑**	**AUROC↑**
Qwen3-VL-8B	Full	8.9	**17.5**	**90.1**	**80.1**
	−Brier	8.5	19.0	87.1	78.0
	−Rank	**4.9**	17.7	88.7	77.2
	BCE only	7.3	18.2	88.7	77.2
LLaVA-NeXT-13B	Full	**5.7**	**18.2**	**87.7**	**78.9**
	−Brier	6.4	18.6	87.2	78.2
	−Rank	10.4	20.9	84.6	72.9
	BCE only	10.5	21.1	84.1	72.2
InternVL3.5-14B	Full	**7.9**	**19.0**	**88.0**	**76.4**
	−Brier	10.9	20.4	87.2	75.2
	−Rank	9.7	19.8	86.6	74.1
	BCE only	10.9	20.4	86.3	73.7
DeepSeek-VL2	Full	**6.0**	**17.9**	**86.1**	**81.1**
	−Brier	7.0	18.1	86.0	80.8
	−Rank	8.4	19.8	83.5	77.0
	BCE only	9.7	20.2	83.6	77.3
Gemma-3-27B	Full	7.0	**19.6**	**85.1**	**76.6**
	−Brier	**6.4**	19.3	85.4	77.0
	−Rank	7.3	19.9	84.2	75.3
	BCE only	7.3	19.7	84.8	75.9

Table 25: Confidence distributions across all 25 runs. "Separation" is the difference between correct and incorrect means; higher indicates better discrimination.

Variant	**Correct**	**Incorrect**	**Separation**
Full (BICR)	70.2	41.6	**28.7**
$-\mathcal{L}_{\text{brier}}$	68.2	39.3	28.9
$-\mathcal{L}_{\text{rank}}$	72.2	49.2	23.0
$\mathcal{L}_{\text{bce}}$ only	71.0	47.3	23.7

Table 26: Full reliability diagram across all 10 bins for each ablation variant, pooled across all 25 runs (5 LVLMs $\times$ 5 seeds). Comparing Full to $-\mathcal{L}_{\text{rank}}$ isolates the contribution of the ranking loss in each bin (with $\mathcal{L}_{\text{brier}}$ held constant); the $\mathcal{L}_{\text{bce}}$-only column shows the joint effect of removing both auxiliary losses. "Gap" = |Pred. − Act.|; lower is better. All values are percentages, with the lowest gap per row bolded.

	Full (BICR)			$-\mathcal{L}_{\text{rank}}$			$\mathcal{L}_{\text{bce}}$ **only**		
Bin	Pred.	Act.	Gap	Pred.	Act.	Gap	Pred.	Act.	Gap
$[0.0, 0.1)$	6.3	23.3	16.9	6.6	20.6	**14.0**	6.7	21.8	15.0
$[0.1, 0.2)$	15.2	31.5	**16.3**	15.4	32.5	17.1	15.4	33.9	18.5
$[0.2, 0.3)$	25.1	38.4	**13.3**	25.2	40.8	15.6	25.3	42.8	17.5
$[0.3, 0.4)$	35.1	45.2	**10.2**	35.2	46.3	11.2	35.1	47.7	12.6
$[0.4, 0.5)$	45.0	50.4	5.4	45.1	49.0	**4.0**	45.0	50.1	5.1
$[0.5, 0.6)$	55.0	56.0	**1.1**	55.0	52.2	2.7	55.0	52.9	2.1
$[0.6, 0.7)$	65.0	63.4	**1.6**	65.0	56.7	8.3	65.0	57.8	7.1
$[0.7, 0.8)$	75.0	72.0	**3.1**	75.0	64.3	10.7	75.0	65.2	9.8
$[0.8, 0.9)$	85.3	83.6	**1.8**	85.3	77.4	7.9	85.3	77.9	7.4
$[0.9, 1.0]$	95.9	94.9	**1.0**	95.9	93.3	2.7	96.0	93.0	3.0

Table 27: Optuna-optimized loss hyperparameters for BICR across LVLMs. Each cell shows the mean across 5 seeds, with the range [min, max] in parentheses. "Arch." reports the most frequently selected `classifier_layers` configuration across the 5 seeds.

LVLM	β **(Brier weight)**	λ **(Rank weight)**	γ **(Margin)**	**Arch.**
Qwen3-VL-8B	0.26 [0.19, 0.39]	0.20 [0.12, 0.24]	0.09 [0.05, 0.17]	(128, 64)
LLaVA-NeXT-13B	0.35 [0.15, 0.49]	0.07 [0.01, 0.25]	0.11 [0.05, 0.14]	(256, 128)
InternVL3.5-14B	0.16 [0.00, 0.40]	0.11 [0.02, 0.22]	0.12 [0.07, 0.19]	(512, 256)
DeepSeek-VL2	0.25 [0.07, 0.48]	0.17 [0.07, 0.25]	0.17 [0.10, 0.23]	(512, 256)
Gemma-3-27B	0.28 [0.22, 0.32]	0.09 [0.02, 0.14]	0.17 [0.10, 0.21]	(256, 128)

Table 28: BICR test-set performance on `Qwen3-VL-8B-Instruct` under five null-image strategies for the blank view. Each cell is mean ± std across 5 seeds; per-seed metrics are computed on all 30,514 pooled test samples. Best value per metric in **bold**.

Null type	**ECE**↓	**BS**↓	**Acc**↑	**F1**↑	**AUCPR**↑	**AUROC**↑
black (current)	**0.0886** ± 0.0167	**0.1747** ± 0.0034	**0.7281** ± 0.0044	**0.7866** ± 0.0068	**0.9014** ± 0.0009	**0.8008** ± 0.0019
white	0.1332 ± 0.0276	0.2223 ± 0.0144	0.6908 ± 0.0026	0.3665 ± 0.0339	0.5125 ± 0.0224	0.6780 ± 0.0383
Gaussian noise	0.1231 ± 0.0346	0.2125 ± 0.0222	0.6961 ± 0.0069	0.3872 ± 0.0636	0.5338 ± 0.0366	0.7035 ± 0.0670
blurred	0.1255 ± 0.0311	0.2153 ± 0.0146	0.6919 ± 0.0052	0.3273 ± 0.0997	0.5231 ± 0.0160	0.7007 ± 0.0277
pixel-shuffled	0.1206 ± 0.0216	0.2101 ± 0.0136	0.6957 ± 0.0069	0.3657 ± 0.0493	0.5354 ± 0.0270	0.7215 ± 0.0292

Table 29: Paired Wilcoxon signed-rank tests against the black baseline ($n = 5$ seeds). † denotes unanimous direction across all five seeds (minimum attainable $p = 0.0625$ for $n = 5$).

Comparison	**ECE**↓	**Brier**↓	**Acc**↑	**F1**↑	**AUCPR**↑	**AUROC**↑
vs. white	0.062 †	0.062 †	0.062 †	0.062 †	0.062 †	0.062 †
vs. Gaussian noise	0.125	0.062 †	0.062 †	0.062 †	0.062 †	0.062 †
vs. blurred	0.188	0.062 †	0.062 †	0.062 †	0.062 †	0.062 †
vs. pixel-shuffled	0.062 †	0.062 †	0.062 †	0.062 †	0.062 †	0.062 †

Table 30: Number of shared test samples per LVLM and dataset. Counts are computed after intersecting hash IDs across all confidence estimation methods for each LVLM. The final row reports the total number of evaluated samples across LVLMs.

LVLM	**GQA**	**POPE**	**LLaVA-Wild**	**MMMU_Pro_4**	**MMMU_Pro_10**	**GMAI-MMBench**	**MME-Finance**	**Total**
Qwen3-VL-8B	12,568	9,000	56	1,711	1,717	4,307	882	30,241
LLaVA-NeXT-13B	12,568	9,000	60	1,713	1,719	4,549	892	30,501
InternVL3.5-14B	12,568	9,000	60	1,713	1,718	4,549	892	30,500
DeepSeek-VL2	12,567	9,000	60	1,656	1,697	4,509	870	30,359
Gemma-3-27B	12,568	9,000	60	1,711	1,711	4,549	892	30,491
Total	**62,839**	**45,000**	**296**	**8,504**	**8,562**	**22,463**	**4,428**	**152,092**

Table 31: Pooled aggregate performance per LVLM. Each cell reports mean ± std across 5 seeds (50 Optuna trials each), where each seed value is the metric computed on the entire shared test subset for that LVLM. Best value per LVLM per metric in **bold**.

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
			`Qwen/Qwen3-VL-8B-Instruct` ($n = 30{,}241$)			
P(True)	43.9	44.2	54.6	67.0	76.2	54.6
Self-Probing	24.4	27.5	69.7	81.2	77.0	59.5
PE	19.8	26.0	67.3	80.4	66.2	52.9
SAPLMA	10.6 ±3.4	19.4 ±0.8	72.8 ±0.5	**82.8** ±0.1	86.4 ±0.9	74.2 ±1.4
PIK	7.5 ±1.8	18.0 ±0.5	72.2 ±0.4	81.5 ±0.6	88.4 ±0.6	77.2 ±0.6
CCPS	28.7 ±2.6	45.6 ±18.9	53.9 ±18.9	54.5 ±32.8	66.2 ±3.5	44.9 ±6.4
II	**5.4** ±1.8	**17.0** ±0.2	**73.3** ±0.7	82.6 ±0.2	89.7 ±0.4	79.6 ±0.6
`BICR (Ours)`	8.9 ±1.5	17.4 ±0.3	72.9 ±0.4	78.9 ±0.6	**90.3** ±0.1	**80.1** ±0.2
			`llava-hf/llava-v1.6-vicuna-13b-hf` ($n = 30{,}501$)			
P(True)	26.2	30.9	45.7	34.8	67.9	54.7
Self-Probing	28.9	30.0	65.2	78.1	81.9	67.3
PE	12.8	23.5	63.0	77.3	77.3	68.5
SAPLMA	16.5 ±2.4	23.0 ±1.1	67.3 ±0.7	**78.8** ±0.2	82.0 ±1.3	72.5 ±1.8
PIK	10.8 ±3.7	19.9 ±1.9	68.9 ±2.7	78.2 ±0.9	87.1 ±1.2	77.3 ±2.4
CCPS	16.3 ±0.9	22.0 ±0.4	67.4 ±0.5	78.8 ±0.1	76.0 ±0.3	72.9 ±0.4
II	13.8 ±1.6	22.3 ±0.7	65.4 ±1.2	77.8 ±0.3	81.4 ±0.8	71.9 ±1.7
`BICR (Ours)`	**5.6** ±1.7	**18.2** ±1.1	**71.4** ±2.2	78.1 ±1.0	**87.7** ±1.3	**78.9** ±2.4
			`OpenGVLab/InternVL3_5-14B-HF` ($n = 30{,}500$)			
P(True)	41.2	41.5	57.2	70.0	78.0	59.5
Self-Probing	21.5	24.7	68.4	80.8	76.8	70.8
PE	16.7	25.9	66.6	80.0	60.3	43.4
SAPLMA	16.6 ±1.2	23.3 ±0.4	69.6 ±0.3	**80.8** ±0.1	76.2 ±0.9	65.4 ±0.8
PIK	10.8 ±2.1	20.2 ±0.8	**70.5** ±0.7	80.3 ±0.5	86.4 ±1.1	73.8 ±1.3
CCPS	14.7 ±1.1	23.8 ±0.4	67.8 ±0.1	80.0 ±0.1	71.6 ±3.0	58.2 ±2.3
II	10.6 ±6.7	21.5 ±1.8	68.8 ±0.5	80.4 ±0.6	82.6 ±1.4	69.3 ±0.7
`BICR (Ours)`	**7.9** ±1.3	**19.0** ±0.5	70.2 ±0.8	77.4 ±1.4	**88.0** ±0.2	**76.4** ±0.3
			`deepseek-ai/deepseek-vl2` ($n = 30{,}359$)			
P(True)	34.1	37.4	48.7	56.2	68.1	52.7
Self-Probing	35.1	37.3	55.2	70.7	74.2	62.2
PE	16.3	24.7	56.4	71.4	72.5	73.5
SAPLMA	12.8 ±3.9	21.5 ±1.4	68.2 ±1.6	**75.2** ±0.4	79.7 ±1.9	77.1 ±1.3
PIK	8.5 ±0.8	19.3 ±0.3	68.8 ±0.8	73.2 ±0.3	84.6 ±0.3	78.5 ±0.3
CCPS	7.7 ±3.7	22.1 ±0.9	62.8 ±2.6	71.0 ±0.6	77.0 ±0.9	70.9 ±1.1
II	7.4 ±2.6	19.0 ±0.8	69.7 ±2.2	74.3 ±0.6	84.5 ±0.4	79.3 ±1.1
`BICR (Ours)`	**6.0** ±0.7	**17.9** ±0.2	**73.6** ±0.8	74.9 ±0.5	**86.2** ±0.4	**81.1** ±0.5
			`google/gemma-3-27b-it` ($n = 30{,}491$)			
P(True)	44.8	45.1	54.2	64.9	74.0	56.8
Self-Probing	27.7	29.5	63.9	76.8	79.2	68.4
PE	26.7	30.3	61.4	76.1	70.1	61.8
SAPLMA	4.6 ±2.5	**19.4** ±0.2	**69.6** ±0.5	77.4 ±0.5	83.8 ±0.4	75.5 ±0.5
PIK	8.9 ±1.0	19.9 ±0.7	68.7 ±0.8	**78.1** ±0.3	85.1 ±0.9	76.1 ±1.2
CCPS	9.0 ±1.4	22.1 ±0.9	67.2 ±0.6	76.0 ±0.4	73.4 ±4.4	68.5 ±2.9
II	**4.3** ±1.6	19.7 ±0.8	67.9 ±1.4	76.9 ±0.6	83.2 ±1.2	74.2 ±2.0
`BICR (Ours)`	7.0 ±1.8	19.6 ±0.4	69.6 ±0.5	75.3 ±0.8	**85.1** ±0.2	**76.6** ±0.5

I.2 Cross-LVLM Pooled Average

Table 32 averages the per-LVLM pooled metrics from Table 31 across all five LVLMs, giving equal weight to each LVLM architecture. BICR achieves the best cross-LVLM average on five of six metrics (ECE 7.1, BS 18.4, ACC 71.5, AUCPR 87.5, AUROC 78.6); SAPLMA edges F1 (79.0 vs. 76.9). The next-best method on the discrimination metrics is P(I Know), which BICR beats by $+1.2$ AUCPR and $+2.0$ AUROC points; on the calibration metrics, the next-best is InternalInspector on ECE (8.3, behind BICR by 1.2 points) and P(I Know) on BS (19.4, behind by 1.0 point).

I.3 Per-LVLM Unweighted Average Across Datasets

The pooled evaluation in §I.1 is dominated by the largest datasets (GQA: 12,568 samples; POPE: 9,000 samples), which together constitute over 70% of the test set. To check whether BICR's advantage holds when each dataset contributes equally, Table 33 reports per-dataset metrics averaged with equal weight, excluding datasets with fewer than 100 shared samples per LVLM (LLaVA-Wild is therefore dropped, leaving six of the seven datasets in the average for every LVLM). On this stricter aggregation, BICR achieves the best BS on every LVLM (five of five), the best ECE on three of five (LLaVA-NeXT-13B, InternVL3.5-14B, DeepSeek-VL2), and the best accuracy on every LVLM (five of five). Discrimination is more contested in this view: BICR leads on AUCPR and AUROC for LLaVA-NeXT-13B, P(I Know) takes both metrics on Qwen3-VL-8B and Gemma-3-27B, P(True) takes both metrics on DeepSeek-VL2, and on InternVL3.5-14B P(True) takes AUCPR while Self-Probing takes AUROC (a known artifact of inference-only methods' near-saturated confidence collapsing into a high-AUC score on the smaller datasets that dominate the unweighted average; see §I.6). The shift relative to the pooled view (§I.1, where BICR led AUCPR and AUROC on every LVLM) decomposes into two distinct effects when we look at the per-dataset breakdown (Table 35). BICR's discrimination lead is largest on GQA and POPE, the two largest datasets, which the pooled view amplifies and the equal-weight view de-emphasizes. On the harder grounding-bound datasets (GMAI-MMBench, MMMU-Pro, MME-Finance), the per-LVLM picture is more nuanced. On GMAI-MMBench, BICR's discrimination gaps to the per-LVLM winner are small (typically 2–4 AUROC points) while its calibration is materially better; on MMMU-Pro, P(True) takes discrimination on several LVLMs, but this reflects the saturation artifact already discussed in §I.6 (inference-only methods collapse to a near-uniform high-confidence score that produces accidentally high AUROC on small datasets) rather than genuine probe-level superiority, and BICR's calibration on MMMU-Pro is the best across LVLMs by a wide margin (cross-LVLM mean ECE 0.149–0.194 vs. next-best InternalInspector 0.270–0.343). This is what the unweighted view exposes: BICR's strongest contribution on the visually-demanding datasets is calibration, not discrimination, and the equal-weight average amplifies these calibration gains, which is why BICR's ECE and BS leads grow under equal weighting (Table 34: ECE gap to InternalInspector grows from 1.2 to 4.2 points, BS gap from 1.5 to 3.7) even as discrimination tightens.

Table 32: Cross-VLM pooled average performance. Each cell is mean ± std of the per-VLM pooled means across all five LVLMs (equal weight per LVLM). Best per metric in **bold**.

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
P(True)	38.0 ±7.0	39.8 ±5.2	52.1 ±4.2	58.6 ±12.8	72.8 ±4.2	55.7 ±2.3
Self-Probing	27.5 ±4.6	29.8 ±4.2	64.5 ±5.1	77.5 ±3.8	77.8 ±2.6	65.6 ±4.1
PE	18.5 ±4.7	26.1 ±2.3	63.0 ±3.9	77.1 ±3.2	69.3 ±5.8	60.0 ±10.8
SAPLMA	12.2 ±4.4	21.3 ±1.7	69.5 ±1.9	**79.0** ±2.6	81.6 ±3.5	72.9 ±4.1
PIK	9.3 ±1.3	19.4 ±0.8	69.8 ±1.3	78.3 ±2.9	86.3 ±1.4	76.6 ±1.6
CCPS	15.3 ±7.5	27.1 ±9.3	63.8 ±5.3	72.1 ±9.3	72.8 ±3.8	63.1 ±10.4
II	8.3 ±3.5	19.9 ±1.9	69.0 ±2.6	78.4 ±2.9	84.3 ±2.9	74.8 ±4.1
BICR (Ours)	**7.1** ±1.2	**18.4** ±0.8	**71.5** ±1.5	76.9 ±1.6	**87.5** ±1.8	**78.6** ±1.9

Table 33: Unweighted average performance per LVLM. For each seed, metrics are first computed per dataset, then averaged with equal weight across datasets with at least 100 shared samples. The reported value is mean ± std of these unweighted averages across 5 seeds. Best per LVLM per metric in **bold**.

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
			`Qwen/Qwen3-VL-8B-Instruct` (6 datasets)			
P(True)	48.4	48.0	50.3	61.3	70.5	65.7
Self-Probing	36.2	38.3	58.5	67.8	65.7	59.0
PE	35.8	37.8	52.8	66.4	54.5	53.0
SAPLMA	20.6 ±3.1	26.2 ±1.7	63.1 ±1.0	**72.5** ±0.3	73.8 ±0.3	69.6 ±0.5
PIK	17.9 ±2.2	24.2 ±1.3	62.6 ±1.0	71.7 ±0.7	**75.3** ±1.1	**71.7** ±1.0
CCPS	35.7 ±5.1	48.5 ±9.8	50.9 ±9.7	51.7 ±22.3	64.8 ±3.0	49.7 ±5.1
II	**9.8** ±2.0	20.8 ±0.4	65.7 ±1.7	71.4 ±1.3	73.7 ±1.2	70.2 ±1.3
BICR (Ours)	10.8 ±0.5	**20.6** ±0.1	**67.7** ±0.4	63.8 ±1.0	74.8 ±0.5	71.3 ±0.5
			`llava-hf/llava-v1.6-vicuna-13b-hf` (6 datasets)			
P(True)	26.6	25.1	58.2	33.8	50.0	63.2
Self-Probing	45.7	44.6	47.3	55.1	50.2	54.4
PE	35.7	33.4	41.7	53.9	46.4	55.8
SAPLMA	32.7 ±3.3	32.4 ±2.4	51.7 ±2.3	**57.3** ±0.2	54.2 ±0.5	64.1 ±0.3
PIK	26.7 ±5.9	27.6 ±4.3	56.1 ±6.7	56.5 ±0.2	58.4 ±0.4	68.5 ±0.7
CCPS	35.3 ±1.0	35.5 ±0.7	48.9 ±0.9	56.7 ±0.2	58.8 ±0.2	65.7 ±0.2
II	30.8 ±2.3	31.7 ±1.6	47.6 ±3.1	55.6 ±0.7	47.0 ±0.5	55.4 ±0.8
BICR (Ours)	**19.4** ±5.8	**22.5** ±3.2	**64.4** ±5.8	55.6 ±1.1	**59.3** ±0.3	**69.4** ±0.3
			`OpenGVLab/InternVL3_5-14B-HF` (6 datasets)			
P(True)	42.4	41.9	55.9	63.9	**71.5**	68.9
Self-Probing	33.8	33.7	56.3	67.0	64.1	**70.7**
PE	35.3	38.0	52.4	65.8	56.9	54.5
SAPLMA	26.5 ±1.2	30.5 ±0.7	60.9 ±0.6	**70.7** ±0.3	63.7 ±0.6	64.5 ±0.5
PIK	25.0 ±2.4	29.1 ±1.7	58.6 ±0.9	69.5 ±0.3	68.9 ±1.4	68.9 ±1.0
CCPS	30.0 ±1.2	34.2 ±0.9	55.3 ±0.1	68.3 ±0.1	60.9 ±0.7	56.9 ±0.8
II	22.6 ±6.6	29.4 ±3.8	57.2 ±1.0	68.7 ±1.3	62.2 ±1.6	59.9 ±1.8
BICR (Ours)	**16.2** ±0.3	**23.3** ±0.2	**63.3** ±0.3	66.5 ±0.9	68.7 ±0.2	69.0 ±0.2
			`deepseek-ai/deepseek-vl2` (6 datasets)			
P(True)	43.0	39.1	40.0	44.2	**52.4**	**65.2**
Self-Probing	50.4	49.8	39.1	49.9	51.3	52.6
PE	31.9	28.3	42.9	50.3	50.1	63.1
SAPLMA	23.7 ±4.1	27.2 ±2.6	59.8 ±3.1	**53.0** ±0.6	48.2 ±0.5	64.7 ±1.1
PIK	19.4 ±1.1	23.3 ±0.9	59.7 ±2.2	47.9 ±1.7	49.5 ±0.3	61.9 ±0.7
CCPS	21.8 ±3.5	23.9 ±2.3	59.1 ±6.9	50.4 ±1.7	46.5 ±1.0	62.5 ±1.5
II	15.3 ±4.2	20.6 ±2.0	66.7 ±5.3	47.0 ±6.0	50.0 ±1.1	65.0 ±0.7
BICR (Ours)	**11.3** ±1.1	**18.4** ±0.7	**74.1** ±2.2	40.0 ±1.9	49.5 ±0.3	62.4 ±0.3
			`google/gemma-3-27b-it` (6 datasets)			
P(True)	44.9	45.0	53.8	60.6	63.8	64.9
Self-Probing	34.1	35.2	55.4	64.8	60.8	62.2
PE	39.1	40.2	49.4	64.2	52.5	55.1
SAPLMA	**11.1** ±1.2	22.6 ±0.4	62.7 ±1.1	64.9 ±1.8	65.7 ±0.3	65.4 ±0.3
PIK	17.3 ±2.2	24.2 ±1.3	60.2 ±2.0	**69.0** ±0.5	**72.2** ±0.9	**71.0** ±0.9
CCPS	20.9 ±3.3	29.4 ±2.6	58.5 ±1.7	63.2 ±1.6	61.8 ±1.1	61.0 ±1.1

(Continued on next page)

(Continued from previous page)

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
II	11.3 ±5.0	23.0 ±1.8	60.5 ±3.3	65.4 ±2.8	64.5 ±1.7	64.7 ±1.3
BICR (Ours)	11.5 ±1.2	**22.1** ±0.3	**64.4** ±1.4	63.7 ±1.6	69.0 ±1.1	68.3 ±0.7

I.4 Cross-LVLM Unweighted Average

Table 34 averages the per-LVLM unweighted metrics from Table 33 across all five LVLMs, giving equal weight to each LVLM architecture and each source dataset. BICR achieves the best calibration (ECE 13.8 vs. next-best II 18.0; BS 21.4 vs. next-best II 25.1) and the best accuracy (66.8 vs. next-best SAPLMA 59.6) by substantial margins. On discrimination, P(I Know) narrowly edges BICR on AUROC (68.4 vs. 68.1) and AUCPR (64.8 vs. 64.2), with both gaps under one point and well within the across-LVLM standard deviations of either method. The trade-off this view exposes is informative: P(I Know) reaches comparable discrimination to BICR but at a substantially worse calibration cost (ECE 21.3 vs. 13.8, a 7.5-point gap; BS 25.7 vs. 21.4, a 4.3-point gap), so the methods that compete with BICR on discrimination do so by being meaningfully more miscalibrated. SAPLMA leads F1 (63.7 vs. BICR 57.9). The headline takeaway from the unweighted view: even when the equal-weight aggregation strips out the dominance of the largest datasets, BICR retains a clear calibration and accuracy advantage and stays within sub-point distance of the strongest baseline on discrimination, while no baseline matches BICR on both axes at once.

I.5 Per-LVLM Per-Dataset Breakdown

Table 35 provides the full per-dataset breakdown for each LVLM. Within each LVLM block, results are grouped by dataset, with methods as rows and metrics as columns. The test set comprises seven source datasets: GQA (visual question answering), GMAI-MMBench (medical imaging), POPE (object hallucination detection), MME-Finance (financial document understanding), MMMU_Pro 4-option and 10-option (multi-choice reasoning), and LLaVA-Wild (open-ended visual dialogue).

The headline pattern is on the multi-choice reasoning datasets where visual grounding is the bottleneck. On MMMU_Pro 10-option BICR wins ECE on four of five LVLMs and accuracy on four of five (with the remaining LVLM on each metric a comfortable second), and the accuracy gap on DeepSeek-VL2 is striking (BICR 84.1 vs. next-best InternalInspector 64.2). On DeepSeek-VL2 specifically, MMMU_Pro 4-opt and 10-opt are extreme low-base-rate regimes (correctness rates of $\sim$11–19%) where BICR's ACC advantage comes from the probe correctly assigning low confidence to the dominant incorrect class; the corresponding AUROC values fall below 0.5 in these cells, indicating that the score's ranking direction is unreliable on the (very few) correct samples even though the calibration gain to incorrect samples is large. We retain the bolding for ACC and ECE on these cells but caution against reading them as joint discrimination wins. MMMU_Pro 4-option shows the same pattern with slightly narrower margins. GMAI-MMBench is more mixed: BICR leads accuracy on four of five LVLMs and ECE on two of five (InternVL and Gemma), but the per-LVLM picture is informative. On Qwen and Gemma, P(I Know) takes both discrimination metrics (AUCPR/AUROC), and on DeepSeek Prompt Ensembles takes both, but in each of these cases the discrimination winner pays a sizable ECE cost relative to BICR: on Gemma, P(I Know) leads on discrimination at ECE 26.2 while BICR sits at comparable discrimination with ECE 11.5; on InternVL, P(True) leads on discrimination (AUCPR 77.8, AUROC 68.7) at ECE 36.8 while BICR reaches comparable AUCPR 75.1 and AUROC 64.6 at ECE 11.3 and BS 24.2 (best on both); on LLaVA, CCPS leads discrimination (AUCPR 48.5, AUROC 63.6) at ECE 30.5, while P(I Know) and BICR achieve nearly identical discrimination (AUCPR 47.7–47.8, AUROC 61.1–61.2) but BICR's ECE (18.2) is materially better than P(I Know)'s (28.5). On the larger near-saturated datasets (GQA, POPE), BICR is competitive but not dominant: P(I Know), SAPLMA, and InternalInspector share calibration leadership on GQA; on POPE BICR matches P(I Know) on AUCPR/AUROC across most LVLMs while P(I Know) and InternalInspector edge calibration. F1 is unfavorable to BICR on several cells because the threshold-based metric rewards the high-recall regime that inference-only baselines and SAPLMA tend to occupy on imbalanced datasets; the threshold-free discrimination metrics (AUCPR, AUROC) tell a more consistent story across all settings.

Table 34: Cross-VLM unweighted average performance. Each cell is mean ± std of the per-VLM unweighted means across all five LVLMs (equal weight per LVLM and per dataset). Best per metric in **bold**.

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
P(True)	41.1 ±7.5	39.8 ±7.9	51.7 ±6.4	52.8 ±11.8	61.6 ±9.0	65.6 ±1.9
Self-Probing	40.0 ±6.7	40.3 ±6.0	51.3 ±7.2	60.9 ±7.2	58.4 ±6.5	59.8 ±6.4
PE	35.6 ±2.3	35.5 ±4.2	47.9 ±4.7	60.1 ±6.7	52.1 ±3.6	56.3 ±3.5
SAPLMA	22.9 ±7.1	27.8 ±3.4	59.6 ±4.1	**63.7** ±7.5	61.1 ±9.0	65.7 ±2.0
PIK	21.3 ±3.8	25.7 ±2.2	59.4 ±2.1	62.9 ±9.2	**64.8** ±9.6	**68.4** ±3.5
CCPS	28.7 ±6.4	34.3 ±8.2	54.5 ±4.0	58.1 ±6.8	58.5 ±6.3	59.2 ±5.5
II	18.0 ±7.8	25.1 ±4.6	59.5 ±6.9	61.6 ±9.0	59.5 ±9.8	63.0 ±5.0
BICR (Ours)	**13.8** ±3.4	**21.4** ±1.7	**66.8** ±3.9	57.9 ±9.7	64.2 ±8.9	68.1 ±3.0

Table 35: Per-VLM per-dataset performance. Each block shows one LVLM with datasets as subgroups. For trained methods, each cell is mean ± std across 5 seeds (50 Optuna trials each). Best value per dataset per metric in **bold**.

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
			`Qwen/Qwen3-VL-8B-Instruct`			
			GMAI-MMBench			
P(True)	45.8	45.7	54.0	**69.6**	68.7	60.6
Self-Probing	33.9	37.6	54.4	68.6	56.7	61.3
PE	31.4	34.7	53.0	69.2	56.2	53.6
SAPLMA	34.2 ±4.2	36.2 ±2.7	53.3 ±0.3	69.3 ±0.1	63.8 ±0.6	59.6 ±0.5
PIK	15.4 ±4.1	24.9 ±1.2	56.9 ±2.7	68.0 ±0.8	**73.2** ±1.2	**68.4** ±1.0
CCPS	34.6 ±16.8	48.8 ±2.3	50.1 ±2.1	47.0 ±25.1	57.0 ±7.0	50.5 ±2.6
II	**10.7** ±4.4	**24.6** ±1.0	57.4 ±1.9	67.9 ±1.6	68.0 ±1.9	65.1 ±1.9
`BICR (Ours)`	13.3 ±5.3	25.3 ±1.3	**59.5** ±1.0	50.1 ±7.4	70.1 ±1.4	64.7 ±1.6
			GQA			
P(True)	36.1	36.6	61.7	73.2	79.4	60.9
Self-Probing	23.6	27.6	70.1	81.7	76.1	54.5
PE	18.1	24.5	69.0	81.6	73.0	54.8
SAPLMA	3.7 ±1.6	**16.3** ±0.2	**75.6** ±0.5	**83.7** ±0.4	**89.3** ±0.2	**79.9** ±0.3
PIK	5.7 ±1.4	17.5 ±0.3	73.0 ±0.8	80.5 ±1.2	88.7 ±0.5	78.1 ±0.5
CCPS	20.8 ±10.4	44.5 ±18.0	54.9 ±17.9	52.1 ±36.2	71.8 ±4.9	51.6 ±3.2
II	**3.6** ±1.0	17.5 ±0.2	73.3 ±0.4	82.2 ±0.2	88.4 ±0.3	77.1 ±0.5
`BICR (Ours)`	11.7 ±1.3	18.6 ±0.4	70.6 ±0.8	76.5 ±1.1	89.2 ±0.2	78.6 ±0.3
			LLaVA-Wild			
P(True)	32.8	31.5	67.9	**71.9**	**85.9**	**82.8**
Self-Probing	45.0	44.5	51.8	65.8	57.8	55.3
PE	38.6	39.8	46.4	63.4	52.3	57.8
SAPLMA	17.2 ±5.5	23.6 ±1.9	62.9 ±3.1	48.3 ±5.7	69.6 ±2.1	66.1 ±2.8
PIK	17.3 ±1.9	21.0 ±1.3	65.7 ±7.9	61.4 ±4.5	76.2 ±2.9	70.9 ±3.7
CCPS	38.1 ±20.0	50.3 ±10.4	48.9 ±10.0	42.1 ±22.9	52.7 ±13.1	45.8 ±15.5
II	**12.8** ±4.3	**20.6** ±1.1	**70.4** ±1.8	67.3 ±3.1	73.9 ±6.1	73.6 ±3.3
`BICR (Ours)`	18.0 ±2.6	21.6 ±1.4	70.0 ±2.6	54.3 ±4.3	79.8 ±2.8	77.9 ±3.6
			MME-Finance			
P(True)	44.3	43.7	54.4	69.2	**82.7**	**79.6**
Self-Probing	43.7	43.8	54.6	**69.3**	74.8	59.1
PE	40.4	40.9	51.6	68.1	59.0	57.9
SAPLMA	10.0 ±4.0	**21.4** ±1.0	**67.7** ±1.5	68.7 ±1.3	78.0 ±0.4	74.2 ±0.6
PIK	16.4 ±3.9	24.8 ±1.3	63.7 ±0.9	68.5 ±3.0	72.0 ±1.4	71.0 ±1.4
CCPS	32.8 ±14.3	46.4 ±7.0	52.8 ±6.8	47.3 ±26.2	58.2 ±8.9	51.9 ±9.9
II	**6.7** ±2.2	22.9 ±0.9	62.1 ±2.3	64.9 ±4.4	69.3 ±2.7	68.0 ±3.0
`BICR (Ours)`	13.7 ±5.1	24.3 ±1.3	62.5 ±1.8	55.7 ±7.5	69.9 ±0.9	70.1 ±1.0
			MMMU_Pro_10			
P(True)	60.4	59.1	37.9	42.2	46.3	**68.6**
Self-Probing	57.1	57.9	38.8	42.2	42.4	52.9
PE	65.0	61.1	23.9	38.6	19.9	42.5
SAPLMA	40.1 ±4.8	39.8 ±3.5	43.2 ±2.6	55.2 ±0.2	52.1 ±0.4	61.4 ±0.4
PIK	41.5 ±4.8	40.0 ±4.3	41.3 ±1.7	54.7 ±0.2	56.5 ±2.6	65.8 ±2.1
CCPS	53.1 ±5.8	57.9 ±3.4	41.5 ±3.1	50.0 ±2.2	54.3 ±5.1	49.5 ±1.7

(Continued on next page)

(Continued from previous page)

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
II	19.7 ±3.9	25.5 ±1.5	56.8 ±5.5	**56.2** ±0.9	56.2 ±3.0	67.2 ±1.5
BICR (Ours)	**12.5** ±2.9	**23.1** ±1.0	**63.2** ±2.8	53.5 ±1.5	**57.5** ±0.9	66.5 ±0.8
MMMU_Pro_4						
P(True)	54.2	53.1	44.1	50.9	53.0	**69.0**
Self-Probing	51.0	52.6	44.4	50.8	48.9	54.3
PE	58.3	55.9	30.7	47.0	25.7	42.0
SAPLMA	32.0 ±5.1	34.7 ±3.0	50.2 ±1.7	**64.3** ±0.3	63.6 ±0.2	62.9 ±0.1
PIK	25.5 ±4.5	30.1 ±2.4	51.8 ±2.0	64.2 ±0.6	65.2 ±2.4	65.8 ±1.9
CCPS	44.3 ±4.8	51.2 ±1.3	48.1 ±1.1	57.6 ±3.7	60.6 ±4.8	49.9 ±0.9
II	15.0 ±3.7	25.4 ±1.0	55.7 ±2.9	63.1 ±1.8	64.7 ±2.1	65.9 ±1.6
BICR (Ours)	**6.3** ±1.2	**23.3** ±0.4	**63.1** ±1.0	54.0 ±2.1	**65.5** ±1.2	66.3 ±1.0
POPE						
P(True)	49.6	49.5	50.1	62.4	93.0	55.4
Self-Probing	7.7	10.2	88.6	94.0	95.6	72.0
PE	**1.4**	9.8	88.7	94.0	93.0	67.2
SAPLMA	3.7 ±1.5	8.7 ±0.3	88.7 ±0.1	**94.0** ±0.0	96.0 ±0.4	79.7 ±2.2
PIK	2.8 ±0.7	**8.2** ±0.2	**88.9** ±0.5	94.0 ±0.3	96.3 ±0.2	81.2 ±1.1
CCPS	28.8 ±22.3	42.2 ±37.9	57.7 ±37.9	56.4 ±46.0	86.9 ±7.4	44.7 ±24.6
II	3.1 ±0.8	8.7 ±0.3	88.6 ±0.1	93.9 ±0.1	95.5 ±0.8	78.0 ±3.5
BICR (Ours)	7.5 ±1.0	9.1 ±0.2	87.4 ±0.4	92.9 ±0.2	**96.4** ±0.2	**81.6** ±0.5
llava-hf/llava-v1.6-vicuna-13b-hf						
GMAI-MMBench						
P(True)	**11.1**	**24.2**	**59.1**	18.5	33.7	49.2
Self-Probing	47.2	47.2	38.1	52.1	29.0	49.5
PE	34.6	35.0	35.3	52.2	34.5	48.3
SAPLMA	43.4 ±4.8	43.5 ±3.7	38.0 ±1.3	51.8 ±0.6	36.9 ±0.6	52.1 ±0.6
PIK	28.5 ±9.8	32.0 ±5.3	47.7 ±8.3	50.9 ±1.4	47.8 ±0.4	61.1 ±0.5
CCPS	30.5 ±2.1	31.2 ±1.3	45.9 ±2.8	**53.8** ±0.4	**48.5** ±0.3	**63.6** ±0.1
II	33.2 ±4.0	35.5 ±1.9	39.0 ±3.2	50.3 ±2.0	36.8 ±1.2	51.6 ±1.3
BICR (Ours)	18.2 ±7.5	26.7 ±3.1	56.9 ±5.9	46.0 ±6.8	47.7 ±0.7	61.2 ±0.6
GQA						
P(True)	37.2	35.2	36.4	22.8	75.0	55.8
Self-Probing	23.4	26.3	70.8	82.7	79.9	55.2
PE	6.9	20.8	70.4	82.6	78.1	59.9
SAPLMA	6.2 ±2.0	17.9 ±0.3	**73.8** ±0.3	83.4 ±0.2	87.4 ±0.4	75.0 ±0.5
PIK	6.7 ±1.2	18.2 ±0.3	71.5 ±0.7	79.8 ±0.9	88.4 ±0.1	75.5 ±0.2
CCPS	5.9 ±0.8	**17.6** ±0.1	73.7 ±0.1	**83.4** ±0.1	**88.5** ±0.0	**76.1** ±0.0
II	**5.1** ±1.1	19.3 ±0.2	71.1 ±0.5	82.0 ±0.7	83.3 ±0.3	68.5 ±0.7
BICR (Ours)	10.5 ±2.5	19.0 ±0.6	69.3 ±1.3	76.6 ±1.8	88.4 ±0.1	75.8 ±0.2
LLaVA-Wild						
P(True)	**8.8**	**18.4**	**73.3**	42.9	**53.9**	71.8
Self-Probing	57.8	54.9	33.3	47.4	42.0	51.1
PE	43.4	38.5	30.0	46.2	51.6	**72.4**
SAPLMA	32.1 ±2.7	31.6 ±1.9	55.0 ±3.7	**53.5** ±3.1	35.5 ±1.3	63.8 ±2.2
PIK	37.5 ±1.7	37.0 ±1.3	45.7 ±2.3	45.4 ±1.1	47.7 ±3.3	58.2 ±1.2
CCPS	43.4 ±0.7	39.9 ±0.9	39.7 ±1.2	46.9 ±0.5	43.1 ±1.8	64.2 ±1.7
II	38.6 ±5.1	36.3 ±3.6	37.0 ±4.8	44.9 ±2.9	37.3 ±5.1	54.5 ±3.3
BICR (Ours)	31.1 ±6.0	30.3 ±4.0	54.0 ±9.0	47.8 ±5.2	48.8 ±3.8	60.6 ±2.8
MME-Finance						
P(True)	**15.6**	**18.2**	**76.3**	35.9	37.6	68.0
Self-Probing	70.0	66.1	23.0	36.0	40.4	58.9
PE	52.3	43.9	21.7	35.7	32.3	59.1
SAPLMA	40.4 ±5.3	34.8 ±4.1	39.9 ±5.6	37.0 ±1.4	31.8 ±1.1	60.9 ±1.7
PIK	45.9 ±2.8	39.2 ±2.4	37.7 ±4.5	37.8 ±0.7	38.5 ±1.0	66.1 ±1.0
CCPS	59.4 ±0.7	53.2 ±0.7	23.9 ±0.3	35.0 ±0.4	35.6 ±1.1	60.2 ±0.5
II	48.7 ±3.5	42.7 ±2.7	28.1 ±5.0	34.9 ±1.1	22.7 ±0.8	50.9 ±0.4
BICR (Ours)	36.5 ±7.2	31.0 ±5.9	49.4 ±8.0	**40.8** ±2.1	**41.8** ±1.4	**69.2** ±0.8
MMMU_Pro_10						
P(True)	**25.2**	**19.6**	**68.0**	**34.5**	24.9	**68.0**
Self-Probing	62.9	59.2	30.5	30.4	28.0	56.0
PE	56.6	45.4	15.8	27.3	19.8	54.2
SAPLMA	54.9 ±4.2	46.4 ±4.0	31.3 ±4.6	33.1 ±0.8	33.8 ±1.2	67.0 ±0.6

(Continued on next page)

(Continued from previous page)

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
PIK	39.6 ±12.8	33.2 ±10.7	46.8 ±16.8	33.4 ±1.7	34.6 ±0.8	63.0 ±1.4
CCPS	58.8 ±1.6	53.1 ±1.5	28.1 ±1.4	31.2 ±0.3	**37.3** ±0.3	55.1 ±0.3
II	51.1 ±3.5	42.7 ±3.1	26.2 ±6.2	30.8 ±0.8	20.0 ±1.0	52.7 ±1.6
BICR (Ours)	25.5 ±11.9	23.3 ±6.2	64.1 ±12.8	34.3 ±1.1	35.8 ±0.3	64.4 ±0.7
MMMU_Pro_4						
P(True)	**22.2**	**19.2**	**69.8**	41.8	34.7	**70.9**
Self-Probing	59.7	56.6	33.7	35.6	30.2	56.5
PE	53.6	44.1	18.9	31.7	21.2	51.4
SAPLMA	45.4 ±4.3	41.1 ±3.2	39.4 ±4.0	**45.1** ±0.6	42.3 ±0.8	65.1 ±0.5
PIK	36.6 ±10.9	34.6 ±8.4	44.7 ±11.7	43.7 ±1.0	44.5 ±0.8	62.7 ±1.3
CCPS	51.7 ±1.5	48.4 ±1.1	34.2 ±1.7	43.4 ±0.4	**46.2** ±0.4	57.5 ±0.3
II	42.1 ±3.7	39.3 ±2.6	33.0 ±4.6	41.9 ±0.8	28.5 ±1.7	51.9 ±1.8
BICR (Ours)	23.0 ±10.2	26.0 ±4.7	59.0 ±9.6	42.5 ±2.0	45.4 ±0.2	63.7 ±0.6
POPE						
P(True)	48.3	34.1	39.8	49.2	94.2	67.1
Self-Probing	10.9	12.0	87.9	93.6	93.8	50.7
PE	10.4	11.2	**88.4**	**93.8**	92.3	61.9
SAPLMA	5.9 ±1.7	10.5 ±0.3	88.0 ±0.4	93.6 ±0.2	92.9 ±0.3	64.7 ±0.8
PIK	2.9 ±1.4	**8.7** ±0.4	88.2 ±0.1	93.7 ±0.1	**96.7** ±0.2	**82.4** ±1.3
CCPS	5.5 ±0.3	9.2 ±0.1	87.9 ±0.3	93.5 ±0.2	96.6 ±0.0	81.6 ±0.3
II	4.5 ±1.3	10.6 ±0.3	88.2 ±0.2	93.7 ±0.1	90.7 ±0.8	56.9 ±3.1
BICR (Ours)	**2.4** ±1.0	8.8 ±0.1	87.6 ±0.4	93.3 ±0.3	96.6 ±0.1	81.9 ±0.4
OpenGVLab/InternVL3_5-14B-HF						
GMAI-MMBench						
P(True)	36.8	37.0	60.5	74.8	**77.8**	**68.7**
Self-Probing	26.6	29.8	60.1	**75.0**	67.3	65.7
PE	23.8	29.4	60.0	75.0	64.9	56.7
SAPLMA	32.2 ±1.5	34.6 ±0.9	60.0 ±0.2	74.6 ±0.3	64.9 ±0.4	55.3 ±0.4
PIK	21.7 ±3.2	27.3 ±1.4	60.5 ±0.6	74.3 ±0.3	76.4 ±0.5	66.8 ±0.4
CCPS	28.7 ±1.4	33.4 ±0.7	57.4 ±0.4	71.6 ±0.5	56.9 ±2.3	46.4 ±2.3
II	17.5 ±10.2	27.8 ±3.4	59.4 ±1.9	72.2 ±3.8	65.9 ±4.1	57.4 ±5.0
BICR (Ours)	**11.3** ±1.5	**24.2** ±0.2	**61.0** ±0.5	68.9 ±1.4	75.1 ±0.4	64.6 ±0.4
GQA						
P(True)	35.9	36.3	62.5	74.9	75.9	58.9
Self-Probing	21.6	25.8	68.3	81.1	71.3	60.3
PE	13.9	23.4	68.4	**81.2**	73.6	55.4
SAPLMA	11.9 ±1.4	21.8 ±0.4	69.4 ±0.3	80.0 ±0.4	79.3 ±0.6	67.4 ±0.6
PIK	**4.8** ±1.1	**18.0** ±0.2	**72.1** ±0.4	80.0 ±0.5	87.8 ±0.3	76.5 ±0.5
CCPS	6.8 ±1.8	20.2 ±0.3	69.5 ±0.3	80.8 ±0.3	82.5 ±0.3	69.1 ±0.6
II	6.1 ±3.2	19.2 ±0.5	70.4 ±1.1	81.0 ±0.7	85.3 ±1.0	72.8 ±1.7
BICR (Ours)	11.1 ±3.9	19.4 ±1.1	69.4 ±1.9	75.5 ±2.8	**88.1** ±0.1	**77.1** ±0.1
LLaVA-Wild						
P(True)	62.3	60.8	36.7	42.4	49.2	**75.1**
Self-Probing	65.2	60.7	25.0	40.0	**59.6**	65.9
PE	57.8	52.0	25.0	40.0	30.6	57.5
SAPLMA	49.5 ±2.8	46.6 ±2.4	43.7 ±2.2	**45.7** ±1.3	32.3 ±2.9	61.2 ±1.9
PIK	46.0 ±3.5	40.4 ±3.4	35.0 ±3.5	39.7 ±1.2	32.4 ±4.2	57.6 ±5.1
CCPS	49.1 ±1.9	43.4 ±2.3	31.3 ±1.9	38.7 ±1.8	38.7 ±2.7	58.2 ±4.1
II	43.0 ±9.6	37.7 ±9.3	34.7 ±12.6	40.3 ±5.6	36.6 ±8.8	60.9 ±7.0
BICR (Ours)	**36.1** ±2.8	**32.9** ±2.2	**44.7** ±4.1	39.9 ±1.4	34.5 ±3.4	54.0 ±3.6
MME-Finance						
P(True)	38.7	37.5	59.0	**71.4**	**83.4**	**81.8**
Self-Probing	40.6	39.4	51.8	68.1	78.3	75.8
PE	38.8	38.7	51.6	68.0	73.2	69.4
SAPLMA	21.9 ±1.3	28.2 ±0.7	59.1 ±0.9	69.6 ±0.7	63.1 ±0.8	65.9 ±0.4
PIK	29.1 ±3.4	30.7 ±2.4	55.1 ±2.4	69.2 ±0.9	67.4 ±1.9	70.3 ±1.2
CCPS	34.7 ±1.2	36.0 ±0.8	52.6 ±0.4	68.3 ±0.2	67.1 ±1.4	64.5 ±1.0
II	16.5 ±8.1	27.8 ±3.7	55.4 ±2.9	66.9 ±1.9	61.1 ±2.9	59.5 ±3.2
BICR (Ours)	**14.5** ±1.9	**24.8** ±0.7	**61.2** ±1.4	69.7 ±0.4	64.5 ±0.7	67.8 ±0.8
MMMU_Pro_10						
P(True)	48.7	47.2	49.9	45.3	**46.2**	**74.7**
Self-Probing	55.9	50.6	33.0	39.6	32.0	67.4

(Continued on next page)

(Continued from previous page)

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
PE	67.6	63.8	22.5	36.7	18.4	41.9
SAPLMA	47.2 ±2.0	45.9 ±1.8	43.1 ±1.8	**50.5** ±0.5	36.5 ±1.2	60.4 ±1.1
PIK	49.3 ±4.4	47.7 ±4.3	35.3 ±2.1	47.1 ±0.4	40.3 ±4.1	56.7 ±3.1
CCPS	56.1 ±2.2	53.5 ±2.4	31.3 ±0.2	46.5 ±0.2	33.0 ±1.3	52.3 ±0.9
II	46.5 ±9.5	46.4 ±8.3	33.4 ±3.1	46.0 ±2.1	29.3 ±5.4	44.9 ±7.2
BICR (Ours)	**28.5** ±1.3	**31.3** ±0.8	**50.1** ±1.4	46.5 ±0.4	42.3 ±0.7	59.3 ±0.5
MMMU_Pro_4						
P(True)	44.6	43.5	53.7	51.6	**55.9**	**77.6**
Self-Probing	51.9	47.2	37.2	45.2	38.6	70.3
PE	63.6	60.6	26.4	41.7	21.7	41.3
SAPLMA	40.4 ±2.0	41.3 ±1.6	48.7 ±1.5	**57.4** ±0.6	43.8 ±1.6	60.2 ±1.4
PIK	39.9 ±4.4	41.7 ±3.5	41.5 ±1.5	53.9 ±0.9	43.6 ±3.4	53.8 ±3.4
CCPS	51.0 ±2.1	49.7 ±1.9	35.4 ±0.1	50.5 ±0.2	35.3 ±1.1	45.0 ±1.3
II	41.1 ±8.7	43.0 ±7.1	39.3 ±2.0	54.2 ±1.1	36.3 ±4.9	46.1 ±6.6
BICR (Ours)	**19.9** ±1.5	**29.0** ±0.6	**54.2** ±1.5	48.6 ±0.9	44.2 ±0.9	55.7 ±0.8
POPE						
P(True)	49.7	49.6	50.1	65.7	90.1	51.7
Self-Probing	6.2	9.7	**87.3**	**93.1**	96.9	84.8
PE	4.0	12.4	85.4	92.1	89.7	62.2
SAPLMA	5.5 ±1.4	11.1 ±0.2	85.1 ±0.3	91.8 ±0.1	94.3 ±0.4	78.0 ±1.1
PIK	5.0 ±2.4	**9.0** ±0.8	87.2 ±1.4	92.6 ±0.8	97.9 ±0.1	89.4 ±0.5
CCPS	**2.6** ±1.0	12.1 ±0.2	85.4 ±0.0	92.1 ±0.0	90.4 ±0.7	64.1 ±2.1
II	8.0 ±3.1	11.8 ±1.1	85.4 ±0.1	92.1 ±0.1	95.1 ±0.9	78.9 ±3.6
BICR (Ours)	12.1 ±1.4	11.0 ±0.6	83.8 ±1.0	89.9 ±0.7	**97.9** ±0.0	**89.6** ±0.1
deepseek-ai/deepseek-vl2						
GMAI-MMBench						
P(True)	35.5	36.1	41.3	54.3	41.4	57.6
Self-Probing	49.3	49.9	37.3	54.0	42.1	49.6
PE	33.8	33.8	37.7	53.9	**49.2**	**61.5**
SAPLMA	33.4 ±5.4	36.2 ±3.4	47.1 ±2.8	**54.4** ±0.2	42.4 ±1.4	58.4 ±0.9
PIK	**8.7** ±2.2	**23.7** ±0.5	60.8 ±2.4	37.1 ±6.9	45.0 ±1.3	58.2 ±1.0
CCPS	23.7 ±5.0	30.2 ±2.3	44.9 ±3.3	51.4 ±1.8	38.2 ±0.5	52.5 ±0.6
II	13.0 ±6.7	25.7 ±2.4	55.7 ±5.5	38.5 ±11.5	42.3 ±2.8	55.0 ±2.8
BICR (Ours)	18.1 ±3.7	26.3 ±1.1	**63.8** ±0.2	13.2 ±5.5	46.2 ±0.7	59.7 ±0.7
GQA						
P(True)	17.8	28.4	56.6	63.2	63.9	58.8
Self-Probing	33.2	38.1	54.3	69.9	65.0	54.0
PE	18.0	26.9	53.8	69.9	66.2	63.6
SAPLMA	6.8 ±4.3	**20.7** ±0.8	**68.0** ±0.8	**72.7** ±1.0	**77.6** ±0.4	**75.5** ±0.2
PIK	**4.3** ±0.5	20.8 ±0.1	67.7 ±0.3	70.1 ±0.7	77.1 ±0.5	74.1 ±0.3
CCPS	7.0 ±2.9	24.4 ±0.6	57.3 ±1.0	66.0 ±2.6	65.3 ±0.3	61.6 ±0.4
II	6.3 ±1.6	21.9 ±0.2	64.9 ±0.6	69.1 ±1.5	75.0 ±0.4	71.4 ±0.4
BICR (Ours)	5.9 ±0.5	20.9 ±0.1	67.7 ±0.2	69.5 ±0.5	77.4 ±0.2	74.4 ±0.2
LLaVA-Wild						
P(True)	39.1	35.3	48.3	**41.5**	48.6	60.1
Self-Probing	61.5	62.2	25.0	40.0	40.6	47.5
PE	31.0	26.7	38.3	41.3	**52.6**	**60.9**
SAPLMA	34.2 ±5.3	32.3 ±4.1	51.0 ±6.1	32.4 ±2.1	37.6 ±3.0	51.4 ±1.8
PIK	35.5 ±4.0	32.8 ±3.0	43.7 ±5.3	34.7 ±1.2	25.0 ±1.1	46.2 ±2.1
CCPS	**17.1** ±3.9	24.3 ±1.4	60.3 ±5.1	24.3 ±6.0	27.4 ±2.2	50.3 ±2.6
II	18.9 ±7.6	**23.5** ±3.6	**62.0** ±9.0	32.7 ±5.2	36.7 ±8.4	56.4 ±8.9
BICR (Ours)	28.9 ±4.3	27.2 ±1.7	58.0 ±2.7	37.6 ±1.4	34.7 ±4.0	52.8 ±1.7
MME-Finance						
P(True)	44.9	39.1	34.3	47.3	55.3	70.6
Self-Probing	59.8	59.3	30.9	45.9	50.1	52.7
PE	35.2	29.6	36.8	48.0	**66.0**	**76.8**
SAPLMA	26.5 ±3.5	28.1 ±2.5	54.4 ±2.8	50.3 ±0.7	46.0 ±0.9	66.9 ±0.9
PIK	18.3 ±1.2	22.2 ±0.7	65.8 ±1.6	**52.3** ±1.8	52.0 ±2.1	70.1 ±2.0
CCPS	16.4 ±3.2	25.3 ±1.5	56.9 ±4.9	31.7 ±5.7	30.2 ±0.9	50.3 ±1.9
II	12.6 ±6.6	22.1 ±1.4	65.7 ±3.5	39.9 ±11.0	43.2 ±2.2	64.8 ±3.1
BICR (Ours)	**10.5** ±1.5	**21.3** ±0.6	**67.5** ±2.0	43.5 ±2.6	45.5 ±1.0	65.4 ±0.4
MMMU_Pro_10						

(Continued on next page)

(Continued from previous page)

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
P(True)	57.4	42.0	27.2	17.7	**28.2**	**74.4**
Self-Probing	74.6	68.9	12.7	16.4	27.8	52.8
PE	49.6	33.2	21.7	16.8	11.3	54.7
SAPLMA	37.7 ±6.3	33.6 ±5.4	51.9 ±7.1	17.9 ±1.5	9.8 ±0.6	49.9 ±3.0
PIK	43.1 ±2.2	31.0 ±2.4	40.4 ±5.8	13.2 ±1.7	9.0 ±0.3	40.2 ±1.9
CCPS	38.9 ±6.9	25.6 ±5.7	55.2 ±17.8	**25.5** ±4.1	19.3 ±2.1	65.9 ±3.1
II	32.5 ±7.6	21.8 ±5.3	64.2 ±15.0	16.7 ±5.9	16.5 ±1.2	56.1 ±3.2
BICR (Ours)	**14.9** ±4.2	**13.5** ±2.6	**84.1** ±5.3	7.6 ±3.4	10.5 ±0.5	43.0 ±0.9
MMMU_Pro_4						
P(True)	53.5	39.5	30.6	24.6	**35.6**	**76.2**
Self-Probing	70.9	66.3	16.2	22.3	30.7	52.2
PE	46.0	32.5	23.3	21.6	16.6	54.4
SAPLMA	34.1 ±5.8	34.5 ±4.5	50.2 ±5.4	29.9 ±2.2	17.4 ±0.9	52.4 ±2.7
PIK	38.1 ±2.4	32.4 ±2.2	37.9 ±4.2	22.8 ±1.9	17.0 ±1.2	42.4 ±1.6
CCPS	31.5 ±6.7	24.5 ±5.0	58.2 ±15.8	**38.5** ±4.7	31.7 ±3.2	67.7 ±3.5
II	25.5 ±6.6	22.4 ±3.6	63.3 ±10.0	25.9 ±8.3	27.0 ±3.0	58.4 ±2.7
BICR (Ours)	**15.6** ±4.2	**18.5** ±2.3	**76.2** ±6.5	14.9 ±2.3	20.3 ±0.8	45.7 ±0.8
POPE						
P(True)	48.9	49.3	50.1	58.4	89.8	53.5
Self-Probing	14.4	16.4	83.1	90.7	91.9	54.2
PE	8.9	13.6	84.1	91.4	91.2	67.3
SAPLMA	3.7 ±2.3	**9.9** ±0.5	**86.8** ±0.7	**92.5** ±0.3	96.0 ±0.3	84.8 ±0.7
PIK	4.0 ±1.8	9.9 ±0.6	85.7 ±0.6	91.8 ±0.3	96.8 ±0.6	86.1 ±1.8
CCPS	13.0 ±3.5	13.6 ±0.9	82.2 ±1.0	89.5 ±0.7	94.1 ±0.3	76.9 ±0.4
II	**2.1** ±0.6	10.0 ±0.5	86.0 ±1.1	92.1 ±0.5	95.9 ±0.7	84.0 ±2.1
BICR (Ours)	2.7 ±0.5	9.9 ±0.2	85.0 ±0.3	91.2 ±0.2	**96.9** ±0.1	**86.1** ±0.5
google/gemma-3-27b-it						
GMAI-MMBench						
P(True)	44.2	44.7	53.7	59.2	59.1	58.3
Self-Probing	30.7	34.3	49.7	65.0	56.3	58.9
PE	36.9	38.5	48.9	**65.7**	51.1	54.1
SAPLMA	20.2 ±2.7	31.2 ±1.4	49.5 ±0.7	55.3 ±3.5	50.0 ±0.2	49.6 ±0.3
PIK	26.2 ±4.4	30.7 ±2.5	50.8 ±1.7	65.5 ±0.3	**66.2** ±0.7	**64.4** ±0.9
CCPS	17.3 ±2.9	29.6 ±1.6	51.0 ±1.8	45.8 ±6.5	50.5 ±2.5	51.3 ±2.8
II	11.7 ±7.6	26.7 ±1.9	52.2 ±2.5	58.3 ±7.5	55.5 ±2.2	56.5 ±1.8
BICR (Ours)	**11.5** ±2.2	**25.3** ±0.6	**56.7** ±1.8	61.8 ±1.8	62.6 ±2.0	62.4 ±1.6
GQA						
P(True)	41.9	42.2	57.1	69.4	68.1	57.9
Self-Probing	33.4	35.1	60.0	74.4	70.1	57.3
PE	28.1	31.6	59.5	74.6	68.3	60.2
SAPLMA	3.7 ±1.4	**18.9** ±0.2	**70.8** ±0.5	**77.6** ±0.3	83.0 ±0.3	**77.6** ±0.4
PIK	**3.0** ±0.6	19.1 ±0.1	70.4 ±0.2	77.0 ±0.3	82.8 ±0.1	76.9 ±0.2
CCPS	4.3 ±1.6	21.1 ±0.2	66.6 ±0.3	74.4 ±0.7	78.2 ±0.2	71.6 ±0.3
II	3.8 ±0.4	20.3 ±0.3	67.7 ±0.8	75.6 ±0.5	80.5 ±0.8	73.8 ±0.9
BICR (Ours)	9.0 ±2.2	20.0 ±0.7	69.5 ±0.9	72.1 ±1.7	**83.2** ±0.2	77.3 ±0.3
LLaVA-Wild						
P(True)	41.0	40.9	58.3	**70.6**	**85.2**	**82.9**
Self-Probing	44.1	42.9	50.0	66.7	75.5	69.1
PE	33.9	35.0	51.7	68.1	74.2	74.0
SAPLMA	15.9 ±3.4	24.7 ±1.5	60.3 ±3.4	64.5 ±3.8	70.9 ±4.1	64.1 ±4.2
PIK	18.2 ±6.1	27.4 ±2.1	52.3 ±5.2	61.9 ±2.4	61.5 ±5.9	57.6 ±5.4
CCPS	**13.4** ±2.8	**23.0** ±0.4	63.0 ±1.6	58.3 ±3.8	71.4 ±1.3	65.5 ±2.9
II	13.7 ±2.7	25.4 ±1.8	57.0 ±6.1	62.0 ±6.3	60.6 ±6.8	58.2 ±5.7
BICR (Ours)	17.8 ±2.5	26.0 ±1.0	**63.7** ±3.2	56.8 ±4.5	68.2 ±1.4	62.9 ±2.3
MME-Finance						
P(True)	49.5	49.2	49.9	**62.7**	**75.0**	**79.3**
Self-Probing	46.7	46.6	47.5	59.5	65.3	62.8
PE	48.7	48.0	42.7	59.9	49.6	55.2
SAPLMA	10.6 ±2.2	**23.3** ±0.6	60.5 ±2.7	59.8 ±2.4	61.6 ±0.8	67.5 ±0.9
PIK	14.5 ±2.1	24.1 ±0.5	59.8 ±1.8	61.1 ±2.3	64.2 ±1.8	68.7 ±2.0
CCPS	19.3 ±3.2	26.3 ±1.7	55.9 ±3.1	59.6 ±1.0	63.0 ±2.8	66.3 ±1.8
II	11.2 ±7.0	24.8 ±3.2	57.6 ±6.6	56.9 ±3.0	56.8 ±6.0	63.5 ±4.8

(Continued on next page)

(Continued from previous page)

Method	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
BICR (Ours)	**10.1** ±2.0	23.8 ±0.6	**63.6** ±1.7	47.1 ±2.4	59.7 ±1.8	64.5 ±1.3
MMMU_Pro_10						
P(True)	43.5	43.5	54.4	51.1	42.5	**70.8**
Self-Probing	43.9	42.8	42.4	45.2	36.3	59.1
PE	60.8	57.4	27.5	43.2	23.4	44.3
SAPLMA	16.7 ±2.6	25.9 ±0.8	53.7 ±3.0	47.7 ±2.5	45.9 ±0.7	58.6 ±0.7
PIK	33.1 ±4.2	32.1 ±2.9	43.9 ±5.2	**54.5** ±0.9	**57.6** ±2.0	68.5 ±1.3
CCPS	45.1 ±11.2	47.2 ±9.4	42.5 ±5.6	48.4 ±3.8	36.9 ±1.5	49.5 ±1.6
II	21.7 ±8.1	27.8 ±4.0	49.2 ±9.2	51.4 ±1.6	45.7 ±2.5	60.0 ±1.6
BICR (Ours)	**16.2** ±4.4	**25.1** ±1.7	**57.4** ±5.3	51.1 ±2.5	51.8 ±2.9	63.3 ±1.6
MMMU_Pro_4						
P(True)	40.2	40.5	57.6	57.3	48.0	**69.5**
Self-Probing	38.4	39.7	47.2	52.8	41.9	59.0
PE	54.4	52.3	33.9	50.6	30.7	47.6
SAPLMA	**9.0** ±2.2	25.0 ±0.4	57.0 ±1.2	58.3 ±2.7	59.9 ±0.3	61.6 ±0.4
PIK	22.7 ±4.4	28.2 ±2.3	51.9 ±3.4	**64.9** ±0.4	**67.5** ±1.3	66.8 ±1.1
CCPS	37.2 ±7.2	41.2 ±5.3	49.5 ±2.3	58.9 ±4.6	49.2 ±1.5	50.6 ±1.8
II	12.1 ±7.8	26.6 ±2.2	52.8 ±2.9	59.6 ±5.1	55.6 ±2.2	57.7 ±2.0
BICR (Ours)	11.0 ±2.0	**24.8** ±0.5	**58.0** ±1.7	61.8 ±2.6	62.8 ±1.2	63.8 ±0.9
POPE						
P(True)	50.1	50.0	49.9	63.8	89.8	53.6
Self-Probing	11.6	12.9	**85.4**	**92.0**	**94.7**	75.8
PE	5.7	13.3	84.1	91.3	91.7	68.8
SAPLMA	6.1 ±0.8	11.5 ±0.3	84.5 ±0.5	90.9 ±0.3	93.7 ±0.3	77.7 ±0.7
PIK	4.4 ±1.0	**11.1** ±0.2	84.5 ±0.5	90.9 ±0.4	94.7 ±0.3	**80.4** ±0.8
CCPS	**2.3** ±0.3	11.1 ±0.0	85.4 ±0.1	91.8 ±0.1	92.9 ±0.4	76.9 ±0.8
II	7.5 ±2.5	11.9 ±0.5	83.8 ±0.6	90.5 ±0.5	92.9 ±0.8	76.9 ±1.7
BICR (Ours)	11.2 ±1.8	13.6 ±0.6	81.4 ±0.5	88.6 ±0.4	93.8 ±0.2	78.5 ±0.3

I.6 Per-LVLM and Per-Dataset Reliability Diagrams

The pooled calibration figure in the main text (§5, Figure 2) collapses all five LVLMs and all seven source datasets into a single curve per method. To check whether BICR's calibration advantage is uniform or concentrates in particular settings, this section splits the same per-bin reliability data along three complementary axes. Figures 5–9 fix one LVLM per figure and split that LVLM's data by source dataset, yielding the finest-grained view (one panel per dataset, all methods overlaid). Figure 10 pools across all datasets to give one panel per LVLM, and Figure 11 pools across all five LVLMs to give one panel per dataset.

How to read each panel. Every panel is a reliability diagram. The dashed diagonal marks perfect calibration. Each method contributes *five translucent curves* (one per seed) drawn in a method-specific color, so the spread of the five curves around their joint mean visualizes seed-to-seed variability for that method. Methods that produce a tight cluster of overlapping curves are seed-stable; methods that produce visibly spread or noisy curves are not. *Line thickness within each curve is proportional to the local bin density*: thicker line segments correspond to confidence ranges where the method placed many test samples (and the curve location is therefore reliable), and thinner segments correspond to sparsely populated bins (where the curve location is dominated by sampling noise). A curve sitting *above* the diagonal indicates that the method is under-confident at that confidence level (it is correct more often than its score implies), while a curve *below* the diagonal indicates over-confidence.

Where BICR's calibration advantage concentrates. Tabulating the best-ECE method on each (LVLM, dataset) cell of the per-LVLM figures (35 cells total: 5 LVLMs $\times$ 7 datasets), BICR achieves the lowest ECE on 13 cells, ranks in the top two on 21 cells, and in the top three on 26 cells of 35. Averaging ECE across the five LVLMs per dataset, BICR is the best-calibrated method on four of the seven datasets: GMAI-MMBench (cross-LVLM mean ECE 0.142, vs. next-best InternalInspector 0.168), MME-Finance (0.166 vs. InternalInspector 0.184), MMMU_Pro 10-option (0.194 vs. InternalInspector 0.343), and MMMU_Pro 4-option (0.149 vs. InternalInspector 0.270). On LLaVA-Wild ($n\approx$56–60 per LVLM) BICR (0.244) is essentially tied with InternalInspector (0.231), and on the two near-saturated datasets, POPE (binary object-presence detection) and GQA (large-scale VQA), internal-state baselines overtake BICR (POPE: P(I Know) 0.029, InternalInspector 0.037, BICR 0.071; GQA: InternalInspector 0.044, P(I Know) 0.046, BICR 0.096). This pattern is consistent with BICR's design intent: enforcing visual contrast through blank-image ranking helps most where visual grounding is the bottleneck (medical imaging, document understanding, multi-choice reasoning), and helps less where the underlying task is so easy that nearly every method already places most probability mass in the rightmost bin.

Visible failure modes of baselines. Three regular patterns are evident across the per-LVLM figures. (i) Inference-only methods (P(True), Self-Probing, Prompt Ensembles) collapse to the high-confidence corner on the harder datasets: their curves on GMAI-MMBench, MMMU_Pro, and MME-Finance crowd into a narrow region close to x=1, reflecting confidence saturation rather than calibration. P(True) is particularly stark: its cross-LVLM mean ECE on POPE is 0.493, an artifact of placing virtually every sample into a single high-confidence bin while accuracy on POPE is around 0.5. (ii) CCPS swings systematically under-then-over the diagonal, producing characteristic S-shaped curves on Qwen3-VL-8B and LLaVA-NeXT-13B and inflating its cross-LVLM mean ECE on the multi-choice datasets to 0.513 (MMMU_Pro 10-option) and 0.445 (MMMU_Pro 4-option), a sign of distortion from its contrastive perturbation step; CCPS also displays the largest visible seed-to-seed spread of any trained method on the multi-choice and LLaVA-Wild panels. (iii) The LLaVA-Wild panel ($n\approx$56–60 per LVLM) is the noisiest in every figure: with so few samples per bin, even well-calibrated methods produce visibly jittered curves and substantial seed-to-seed disagreement; the panel should therefore be read as bounded by small-sample variance rather than as a clean miscalibration signal.

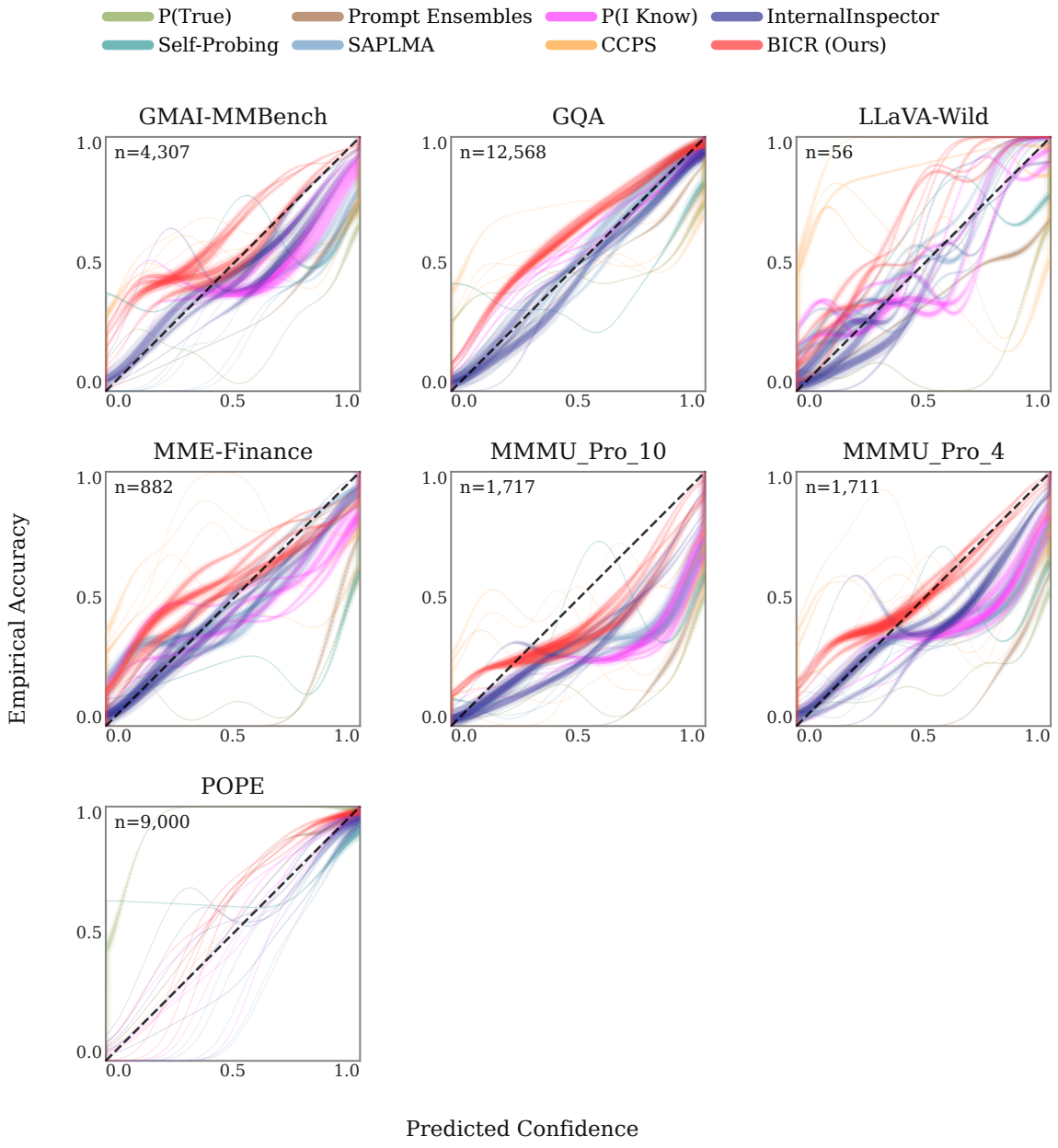


Figure 5: Per-dataset reliability diagrams for `Qwen/Qwen3-VL-8B-Instruct`. Each panel corresponds to one of the seven source datasets (n in panel titles denotes the number of shared test samples for this LVLM). Each method contributes five translucent curves in a method-specific color, one per seed, so the spread visualizes seed-to-seed variability. Visualization style influenced by reliability diagrams in Nakkiran et al. (2025).

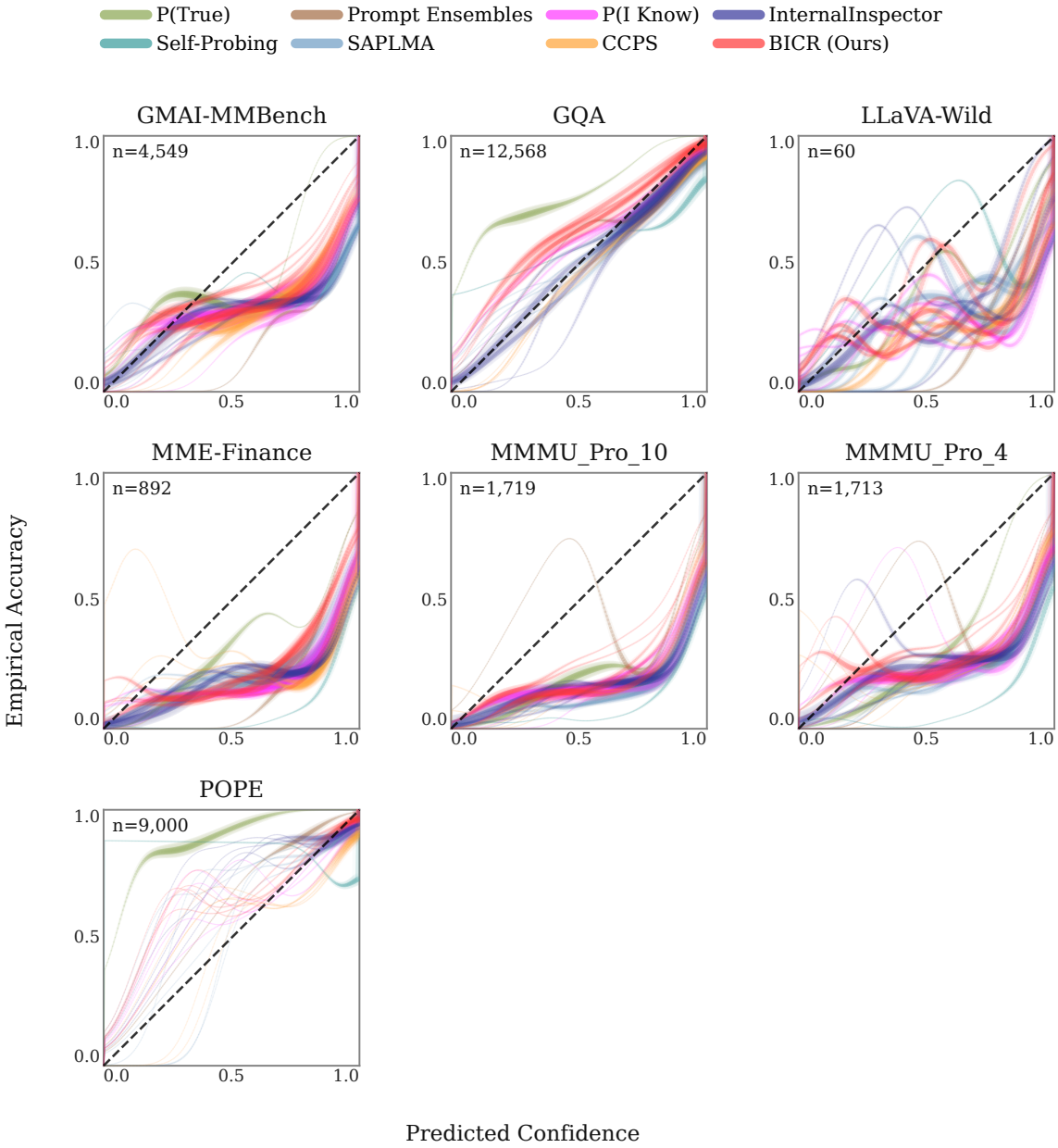


Figure 6: Per-dataset reliability diagrams for `llava-hf/llava-v1.6-vicuna-13b-hf`. Plotting conventions match Figure 5.

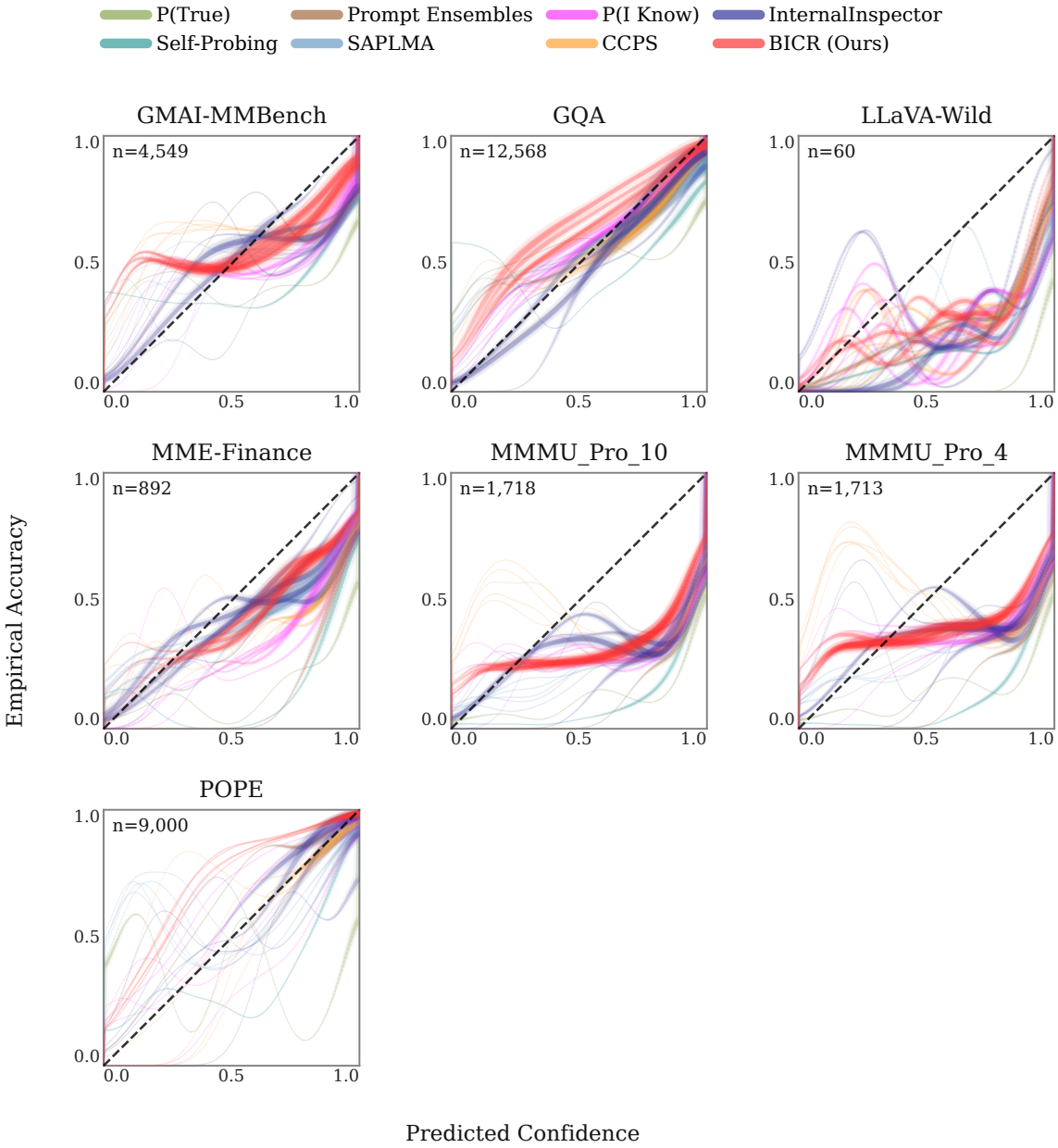


Figure 7: Per-dataset reliability diagrams for `OpenGVLab/InternVL3_5-14B-HF`. Plotting conventions match Figure 5.

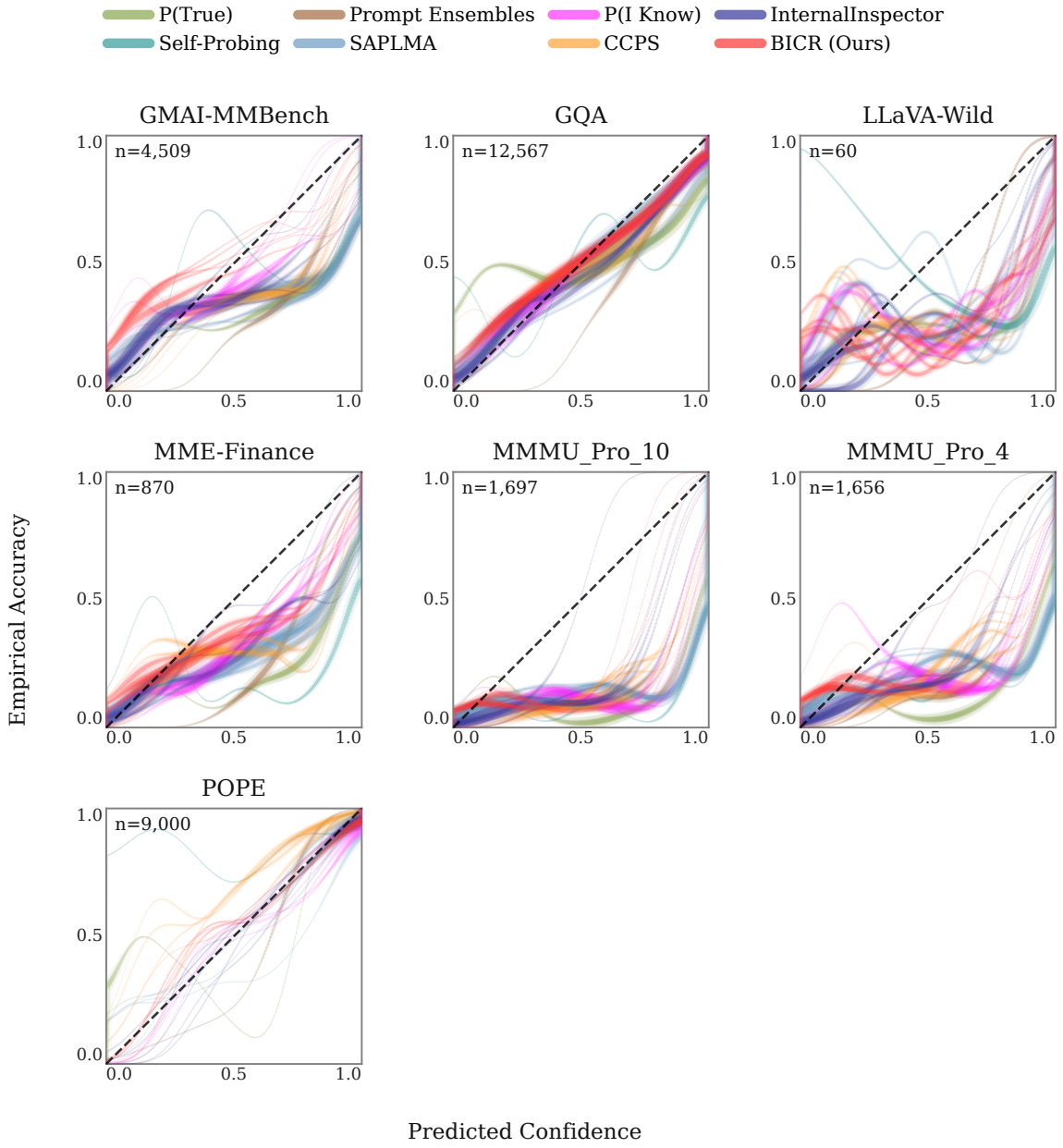


Figure 8: Per-dataset reliability diagrams for `deepseek-ai/deepseek-vl2`. Plotting conventions match Figure 5.

Pooled views, fixed-LVLM slice. Figure 10 fixes the LVLM and pools every (ground truth, confidence) pair across all seven source datasets, yielding one panel per LVLM. The five LVLM subsets are within $\sim$1% of each other in size (n=30,241 to n=30,501), so visual differences across panels reflect LVLM behaviour rather than sample-size artifacts. Two patterns dominate. First, the inference-only baselines fail in characteristic ways: P(True) sits as a tight cluster of curves near x=1 but at empirical accuracies of only $\sim$0.5–0.7 on Qwen3-VL-8B, LLaVA-NeXT-13B, and InternVL3.5-14B (severe high-confidence saturation); Self-Probing produces visibly spread per-seed curves with chaotic low-confidence excursions on every backbone except Gemma-3-27B; CCPS traces an S-shape on Qwen3-VL-8B and a flat over-confident plateau on DeepSeek-VL2. By contrast, `BICR` tracks the diagonal closely on all five backbones with a tight per-seed cluster, with no panel where it collapses to the corner or swings systematically away from the diagonal, indicating that its calibration profile is consistent across model families rather than tuned to one LVLM. Second, Gemma-3-27B is visibly the easiest LVLM to calibrate: nearly every method sits on or near the diagonal in that panel, suggesting the harder calibration cases are upstream in the weaker backbones rather than inherent to the methods.

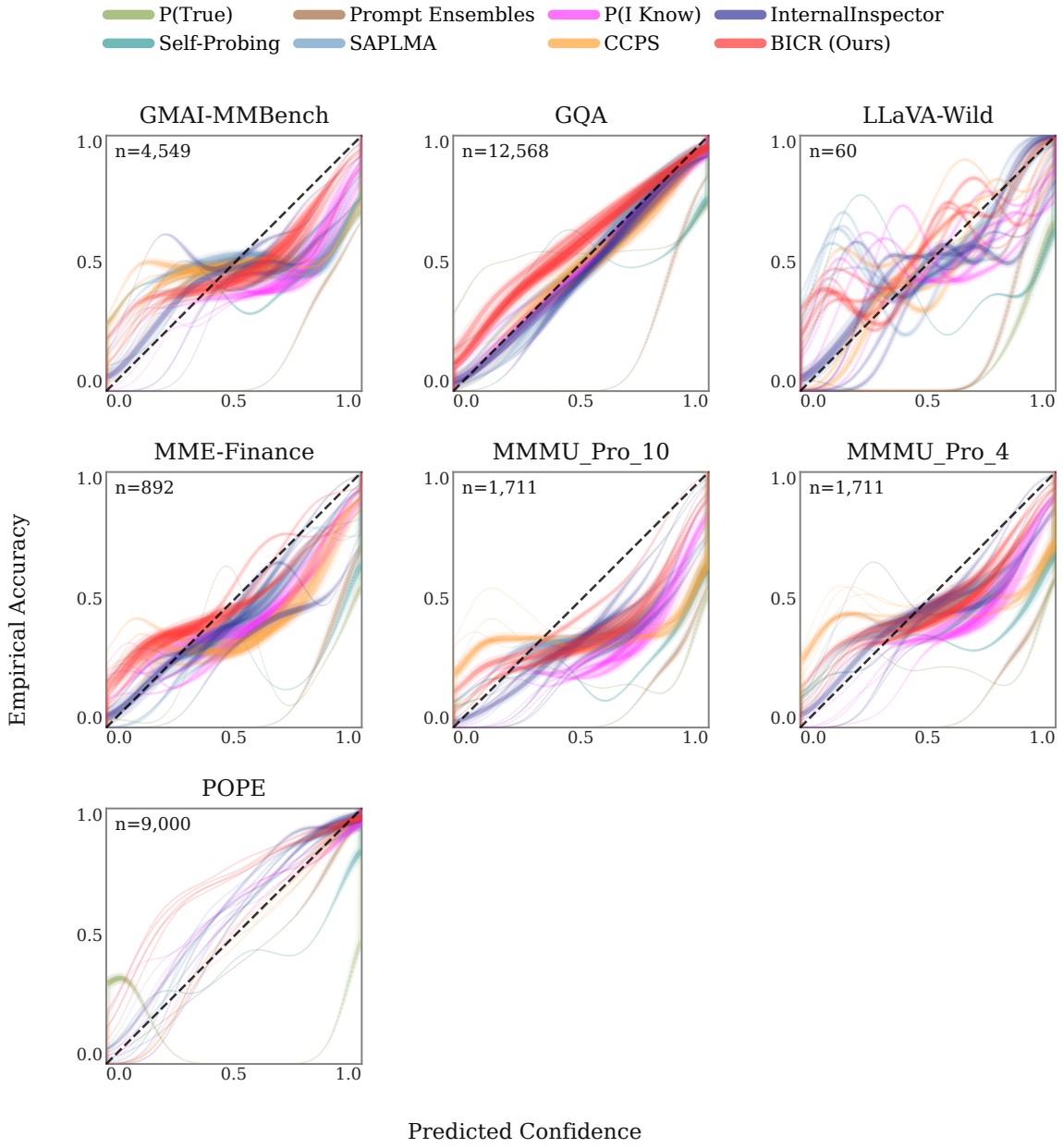


Figure 9: Per-dataset reliability diagrams for `google/gemma-3-27b-it`. Plotting conventions match Figure 5.

Pooled views, fixed-dataset slice. Figure 11 reverses the slicing: it fixes the source dataset and pools across all five LVLMs, with sample counts ranging from n=296 on LLaVA-Wild (visibly the only panel where bin noise dominates) to n=62,839 on GQA. The four hardest grounding-bound datasets (GMAI-MMBench, MME-Finance, MMMU_Pro 4-option, MMMU_Pro 10-option) drive a uniform pattern in which *every* method sits well below the diagonal, i.e., every method is overconfident on these datasets; what differs is the size of the gap, and BICR's curves are consistently the closest to the diagonal from below in those panels. On POPE the pattern inverts: predictions are pushed into the upper-right corner where most curves are mildly under-confident, and the simpler internal-state baselines that excel on near-saturated tasks (P(I Know), InternalInspector) sit closer to the diagonal than BICR. The two pooled figures together make the trade-off in Section 5 concrete: BICR's advantage is a calibration improvement on tasks where visual grounding is the bottleneck and overconfidence is universal, not a uniform improvement on every regime.

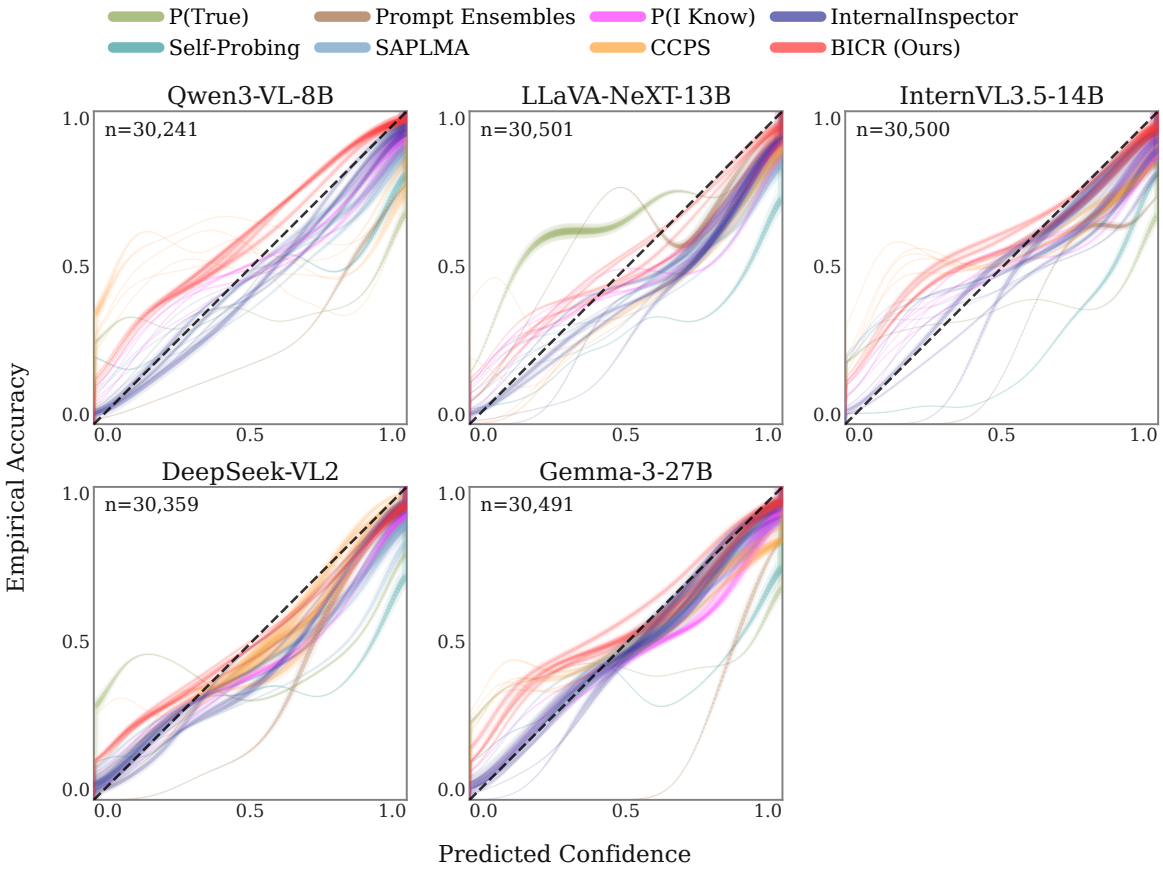


Figure 10: Pooled per-LVLM reliability diagrams. Each panel fixes one LVLM and pools every (ground truth, confidence) pair across all seven source datasets, yielding one panel per LVLM (n in panel titles is the size of the shared test subset for that LVLM). Each method contributes five translucent curves in a method-specific color, one per seed; line thickness within each curve is proportional to local bin density.

I.7 Loss Component Ablation

Table 36 reports the cross-LVLM average metrics for the full BICR model and three ablation variants across all six reported metrics. The full model wins every metric. Removing $\mathcal{L}_{\text{rank}}$ accounts for the largest discrimination drop (-2.0 AUCPR, -3.3 AUROC, -2.8 ACC), confirming that the blank-image ranking signal is the principal driver of BICR's discriminative gain. Removing $\mathcal{L}_{\text{brier}}$ produces a smaller but consistent calibration regression ($+1.4$ ECE, $+0.6$ BS) with negligible discrimination cost, identifying the Brier term as a calibration-specific contribution. The $\mathcal{L}_{\text{bce}}$-only configuration is the worst (or tied with $-\mathcal{L}_{\text{rank}}$ on AUROC) on every metric and behaves as the lower bound for what a single-objective probe can achieve in our setting. A more detailed analysis of these ablations, including per-LVLM breakdowns, statistical significance tests, and behavioral effects on confidence distributions and calibration shape, is provided in Appendix H.

I.8 Statistical Significance

Pooled aggregation. Table 37 reports paired Wilcoxon signed-rank test p-values comparing BICR against each trained baseline under pooled aggregation across 25 (LVLM, seed) observations. Inference-only methods (P(True), Self-Probing, Prompt Ensembles) are excluded since their cross-LVLM AUROC gaps to BICR are sizable (Table 31), making formal significance testing against trained baselines the more informative comparison. BICR's improvements on AUCPR and AUROC are highly significant ($p < 0.001$) against every trained baseline. Calibration improvements are significant against P(I Know), SAPLMA, and CCPS, but BICR is statistically indistinguishable from InternalInspector on ECE ($p = 0.525$) under this aggregation. On BS, BICR significantly improves over InternalInspector ($p < 0.01$) despite the ECE tie.

Equal-weight per-dataset aggregation. Table 38 reports the same significance tests but computed on the per-LVLM unweighted-across-datasets metrics from Table 33, again across 25 (LVLM, seed) observations. Two patterns differ from the pooled view in instructive ways. First, BICR's calibration tie with InternalInspector disappears under equal weighting: BICR significantly beats II on every metric including ECE ($p = 0.007$), reflecting that II's calibration parity with BICR under pooled aggregation came from its strong performance on GQA and POPE rather than uniform calibration across datasets. Second, BICR is statistically indistinguishable from P(I Know) on discrimination under this aggregation (AUCPR $p = 0.381$, AUROC $p = 0.615$), formalizing the trade-off discussed in §I.4: P(I Know) reaches comparable discrimination to BICR on the equal-weight view but at significantly worse calibration ($p < 0.001$ on both ECE and BS), so no baseline matches BICR on both axes simultaneously.

Cluster-aware significance. Both Wilcoxon analyses above treat each (LVLM, seed) tuple as an independent paired observation, but the 5 seeds within a given LVLM share the same frozen weights and the same test set, so the truly independent unit of analysis is the LVLM (n=5). To verify our headline conclusions hold under this stricter independence assumption, Table 39 reports a cluster bootstrap (10,000 resamples) over LVLM-level seed-means with Holm-Bonferroni correction across the 4 metrics within each comparison. BICR's discrimination advantages are robust to this stricter test: AUCPR and AUROC improvements are highly significant ($p < 0.001$) against every trained baseline. Calibration improvements remain significant against P(I Know), SAPLMA, and CCPS (cluster-bootstrap $p < 0.05$ on ECE, $p < 0.001$ on BS), strengthening the n=25 claims under a more conservative test. Against InternalInspector, the cluster-bootstrap evidence on calibration is weaker than under n=25 (ECE n.s., BS marginal at $p = 0.075$ Holm-corrected), consistent with our existing observation that BICR's calibration parity with II under pooled aggregation is largely driven by II's strong performance on GQA and POPE rather than a uniform across-LVLM effect; the discrimination advantage over II remains highly significant on both AUCPR and AUROC ($p < 0.001$). Together with the equal-weight evidence in Table 38, the cluster-aware analysis confirms that no baseline matches BICR on both calibration and discrimination simultaneously under any aggregation protocol.

Together, the three views support the joint-axis framing of the paper: BICR is the only method that is statistically at least as good as the next-best baseline on *both* calibration and discrimination simultaneously across pooled, equal-weight, and cluster-aware analyses, and the specific runner-ups shift between aggregations (InternalInspector under pooled, P(I Know) under unweighted) without

Table 36: Loss component ablation for BICR, averaged across 5 LVLMs and 5 seeds (25 runs). Each row removes one or more auxiliary loss terms. Metrics are computed on the shared test subset, averaged across seeds within each LVLM, then mean ± std across LVLMs. Best per metric in **bold**.

Variant	ECE↓	BS↓	ACC↑	F1↑	AUCPR↑	AUROC↑
Full (BICR)	**7.1** ±1.2	**18.4** ±0.8	**71.5** ±1.5	**76.9** ±1.6	**87.5** ±1.8	**78.6** ±1.9
$-\mathcal{L}_{\text{brier}}$	8.5 ±2.4	19.0 ±0.8	70.6 ±1.5	75.5 ±0.9	87.1 ±1.5	78.0 ±1.9
$-\mathcal{L}_{\text{rank}}$	8.1 ±1.9	19.6 ±1.1	68.7 ±1.9	75.8 ±2.6	85.6 ±1.9	75.3 ±1.7
$\mathcal{L}_{\text{bce}}$ only	9.1 ±1.5	19.9 ±1.0	68.3 ±1.8	75.0 ±2.0	85.5 ±1.9	75.3 ±2.1

Table 37: Statistical significance of BICR vs. trained baselines under pooled aggregation (paired Wilcoxon signed-rank test, n=25). Inference-only methods are excluded as their performance gaps exceed 10 AUROC points on every LVLM. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; "n.s." denotes $p \geq 0.05$.

Comparison	ECE	BS	AUCPR	AUROC
vs PIK	<0.01**	<0.001***	<0.001***	<0.001***
vs SAPLMA	<0.001***	<0.001***	<0.001***	<0.001***
vs II	n.s.	<0.01**	<0.001***	<0.001***
vs CCPS	<0.001***	<0.001***	<0.001***	<0.001***

any single baseline matching BICR on both axes at once.

J Direct Behavioral Test: Calibration on Image-Invariant Samples

The aggregate metrics in §5 and Appendix I establish that BICR dominates calibration and discrimination at the benchmark level, but they do not, on their own, separate the mechanism we claim (BICR's rank loss suppresses confidence on predictions whose internal representation is largely determined by the language prior rather than by the image) from the alternative that BICR is a mild regulariser whose gains are spread evenly across the test set. The two stories produce identical aggregate numbers; they make different predictions on a specific sub-population of the test set. This appendix isolates that sub-population and reports the per-method calibration on it.

J.1 Sub-population: Behaviorally Image-Invariant Samples

For every test sample we run the LVLM on two forward passes: one with the original (image, question) pair, and one in which the original image is replaced by a random natural image drawn uniformly from the union of training-split images across all seven source datasets, with the question and prompt held fixed and the source sample's own image excluded. The substitute image is paired to the sample by a deterministic seed (stable_hash(sample_id ⊕ 'swap') ⊕ 42, where stable_hash takes the first four bytes of the SHA-256 hash), so the substitution is reproducible and identical across LVLMs and across confidence-estimation methods. From each pair of forward passes we record two diagnostics that summarise how much the LVLM's behaviour changed under image substitution:

- $\text{flip}_{\text{swap}} \in \{0, 1\}$: whether the LVLM's first generated token differs between the real-image pass and the image-substituted pass. The first generated token under each image condition is the argmax over the next-token logit distribution at the prompt's final unmasked position rather than a sample from a generation loop. A value of 0 means the LVLM produced the same first token in both conditions and is therefore generating its answer in a way that does not, at the level of the immediate next-token decision, depend on which image is in front of it.

- $\text{dp}_{\text{swap}} \in [0, 1]$: the drop in probability of the real-image top-1 token under the image-substituted pass, computed at the same prompt position. A value near zero means image substitution barely perturbs the LVLM's output distribution; a value near one means the substituted image makes the original top-1 token essentially impossible.

These signals are properties of the frozen LVLM, not of any confidence estimator: they depend only on what tokens the LVLM generates under the two image conditions. None of the eight methods we benchmark, including BICR, uses either signal as input. Stratifying the test set by $\text{flip}_{\text{swap}}$ or dp_{swap} is

Table 38: Statistical significance of BICR vs. trained baselines under per-dataset equal-weight aggregation (paired Wilcoxon signed-rank test, n=25). Same conventions as Table 37.

Comparison	ECE	BS	AUCPR	AUROC
vs PIK	$<0.001^{***}$	$<0.001^{***}$	n.s.	n.s.
vs SAPLMA	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
vs II	0.007^{**}	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
vs CCPS	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$

Table 39: Cluster-aware significance of BICR vs trained baselines under pooled aggregation. Each row reports the mean per-LVLM delta (BICR minus baseline) across 5 LVLMs, with significance assessed by a cluster bootstrap (10,000 resamples) over LVLM-level seed-means and Holm-Bonferroni correction across the 4 metrics within each comparison. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; n.s. denotes $p \geq 0.05$.

Comparison	ECE ↓		BS ↓		AUCPR ↑		AUROC ↑	
	Mean Δ	p	Mean Δ	p	Mean Δ	p	Mean Δ	p
BICR vs P(IK)	-0.0222	0.025^{*}	-0.0101	$<0.001^{***}$	+0.0117	$<0.001^{***}$	+0.0206	$<0.001^{***}$
BICR vs SAPLMA	-0.0513	0.016^{*}	-0.0291	$<0.001^{***}$	+0.0584	$<0.001^{***}$	+0.0569	$<0.001^{***}$
BICR vs I^2	-0.0121	n.s.	-0.0146	n.s.	+0.0316	$<0.001^{***}$	+0.0378	$<0.001^{***}$
BICR vs CCPS	-0.0818	$<0.001^{***}$	-0.0873	$<0.001^{***}$	+0.1463	$<0.001^{***}$	+0.1553	$<0.001^{***}$

therefore a behavioural diagnostic computed from a different intervention than the one any method we evaluate sees at training time; in particular, BICR's rank loss contrasts the real-image hidden state against a blank-image hidden state and never sees the image-substituted pass.

We refer to the sub-population $\text{flip}_{\text{swap}} = 0$ as the *image-invariant subset*: samples on which the LVLM's first-token decision is the same with the real image as with a random natural one. By behavior, the LVLM on these samples is producing its answer from the language prior (or from token-level cues in the prompt) rather than from the visual content. If BICR's rank loss is genuinely teaching the probe to assign lower confidence when the prediction is not anchored in the image, the image-invariant subset is exactly where the effect should be visible. Per-LVLM sample counts for the image-invariant subset and for the failure subset defined in §J.2 are reported in Table 40; the image-invariant subset spans 30.3% of the shared-test population on Gemma-3-27B (the LVLM least sensitive to image substitution) to 62.9% on DeepSeek-VL2 (the most sensitive). The shared-test totals here are slightly smaller than those in Table 30 because this analysis additionally requires the swap-view diagnostic to be present, which excludes a small number of samples on which the swap-view forward pass failed.

Table 40: Per-LVLM sample counts for the image-invariant subset ($\text{flip}_{\text{swap}} = 0$; §J.1) and for the failure subset (image-invariant, incorrect under the main-paper correctness labels, and predicted with original-top-1 probability > 0.8). Percentages are over each LVLM's shared-test total under this analysis.

LVLM	N_{A1}	$\%_{\text{A1}}$	N_{C3}	$\%_{\text{C3}}$
Qwen3-VL-8B	11,268	37.4	2,299	7.6
LLaVA-1.6-13B	13,457	46.8	1,121	3.9
InternVL3.5-14B	15,544	51.4	3,346	11.1
Gemma3-27B	8,980	30.3	2,308	7.8
DeepSeek-VL2	18,679	62.9	4,108	13.8

J.2 Methodology

Subsets analyzed. We focus on three views of the test population. The first is the *image-invariant subset* defined above: all samples with $\text{flip}_{\text{swap}} = 0$, pooled across the seven source datasets and the five LVLMs. This is the population the rank loss is most directly designed to address. The second is the *failure subset*: the image-invariant subset further restricted to samples on which the LVLM was both incorrect (under the same per-sample correctness labels driving the main-paper accuracy results) and assigned the original top-1 token a probability above 0.8. The 0.8 threshold is fixed a priori and not tuned. The failure subset isolates the high-stakes cases where the model was confident, the model was wrong, and the model was not using the image, which is the population on which the grounding-detection mechanism is most directly tested. The third is a *cross-subset summary* com-

puted over twelve different ways of identifying samples that the LVLM treats as image-invariant; we describe its construction in §J.5.

Methods compared. We compare BICR against the same seven baselines benchmarked throughout the paper: P(True), Self-Probing, Prompt Ensemble, P(I Know), SAPLMA, InternalInspector, and CCPS. For trainable methods we average per-sample $P(\text{correct})$ scores across the five seeds before any subset metric is computed; inference-only methods contribute their single test-time score. All eight methods are evaluated on the same shared test subset described in Table 40.

Significance test. For the image-invariant subset we run a paired bootstrap on the per-sample BS difference, restricted to the subset itself: for each of $B = 2{,}000$ resamples (RNG seed 23) we draw n samples with replacement (where n is the size of the subset), paired by sample identifier, and compute $\mathrm{BS}^{\mathrm{baseline}} - \mathrm{BS}^{\mathrm{BICR}}$, where per-sample BS is $(p_{\mathrm{correct}} - y)^2$. Positive values mean BICR achieves lower BS (better calibration) than the baseline on the image-invariant population. We report the mean over resamples and the 95% percentile interval; an interval strictly above zero implies BICR is significantly better calibrated than the baseline on this population. We choose BS on the subset rather than the perhaps-more-obvious mean-confidence gap between the subset and its complement because the latter rewards methods that simply shift their confidence down on hard samples regardless of correctness, an axis on which one of the baselines wins by being uniformly under-confident rather than by detecting grounding.

J.3 Headline Result on the Image-Invariant Subset

Table 41 reports the per-LVLM, per-method metrics on the image-invariant subset. Rows are grouped by LVLM and split by metric; bold marks the best method per row. On the four metrics that require the confidence to be accurate on this subset rather than merely low, BICR achieves the best per-LVLM value on BS in four of five LVLMs (the exception is Gemma, where SAPLMA wins by 0.011 BS), on ECE in three of five, on AUCPR in four of five, and on AUROC in three of five. On Mean Conf, P(True) wins on four of five LVLMs by a route we discuss next; on InternVL, where P(True) does not produce its lowest-confidence regime, BICR's mean confidence is the lowest of any method.

Low Mean Conf is not the same as good calibration. P(True) attains low mean confidence on the image-invariant subset across most LVLMs, well below BICR's. This reflects P(True)'s confidence distribution rather than grounding sensitivity: P(True) assigns near-uniform low confidence to nearly every sample regardless of correctness, which the same table shows directly in P(True)'s BS of 0.35–0.61 and ECE of 0.30–0.61 on the same subset. By the metrics that require confidence to track accuracy, P(True) is the worst method we evaluate on this subset, not the best. The same caveat applies to any method whose mean is low because it is uniformly under-confident; the calibration-faithful metrics expose the difference between low-because-accurate and low-because-degenerate.

Significance test on BS. Table 42 reports the paired-bootstrap mean and 95% confidence interval of $(\mathrm{BS}^{\mathrm{baseline}} - \mathrm{BS}^{\mathrm{BICR}})$ on the image-invariant subset. Positive values indicate BICR is better calibrated than the baseline; bold marks intervals strictly above zero. BICR's BS is significantly lower than the baseline's on 31 of 35 (LVLM, baseline) pairs. The four exceptions are all on Gemma-3-27B (small negative differences against InternalInspector, P(I Know), and SAPLMA; an interval straddling zero against CCPS), consistent with Gemma being BICR's weakest LVLM in the main results (§I.1). Per-cell intervals are uncorrected for multiple testing across the 35 pairs; a small number of marginal cells with mean differences near $+0.005$ are sensitive to standard correction, but the cells against P(True), Prompt Ensemble, and Self-Probing (mean differences in the $+0.06$ to $+0.47$ range) and the majority of cells against the trained baselines (mean differences in the $+0.01$ to $+0.10$ range) are well clear of any reasonable correction. The largest BICR margins are against P(True), at $+0.18$ to $+0.47$ across LVLMs, quantifying the gap between low-because-accurate and low-because-degenerate: by the calibration-faithful test, P(True) is significantly worse than BICR on every LVLM by a wide margin.

J.4 Failure Subset (Image-Invariant, Confidently Wrong)

The image-invariant subset also contains a benign sub-class: samples for which the model produces

Table 41: Per-method calibration and discrimination metrics on the image-invariant subset, the sub-population on which the LVLM's argmax next-token does not change when the real image is replaced by a random natural one (§J.1). Lower is better for Mean Conf, BS, and ECE; higher is better for AUCPR and AUROC. Bold marks the best method per (LVLM, metric) row.

LVLM	Metric	BICR	II	PIK	SAPLMA	CCPS	PE	P(True)	Self-Probing
Qwen3-VL-8B	Conf	0.690	0.810	0.805	0.836	0.616	0.875	**0.397**	0.941
	Brier	**0.143**	0.148	0.157	0.173	0.239	0.212	0.612	0.237
	ECE	**0.045**	0.075	0.076	0.104	0.233	0.139	0.611	0.215
	AUCPR	**0.932**	0.929	0.921	0.905	0.636	0.755	0.700	0.781
	AUROC	0.835	**0.835**	0.807	0.776	0.351	0.585	0.376	0.605
LLaVA-1.6-13B	Conf	0.689	0.801	0.766	0.826	0.845	0.761	**0.340**	0.919
	Brier	**0.169**	0.218	0.191	0.225	0.233	0.221	0.349	0.285
	ECE	**0.051**	0.159	0.126	0.185	0.204	0.100	0.302	0.275
	AUCPR	**0.897**	0.857	0.891	0.846	0.742	0.793	0.611	0.753
	AUROC	**0.818**	0.775	0.806	0.759	0.727	0.680	0.441	0.683
InternVL3.5-14B	Conf	**0.625**	0.767	0.744	0.795	0.794	0.835	0.710	0.866
	Brier	**0.187**	0.202	0.198	0.220	0.234	0.254	0.448	0.232
	ECE	**0.067**	0.104	0.094	0.142	0.136	0.163	0.445	0.198
	AUCPR	**0.880**	0.857	0.865	0.782	0.725	0.620	0.752	0.791
	AUROC	**0.769**	0.735	0.745	0.687	0.604	0.452	0.549	0.717
Gemma3-27B	Conf	0.663	0.725	0.756	0.732	0.745	0.888	**0.501**	0.935
	Brier	0.164	0.158	0.158	**0.153**	0.167	0.228	0.565	0.242
	ECE	0.070	0.036	0.048	**0.029**	0.036	0.175	0.563	0.232
	AUCPR	0.902	0.901	**0.907**	0.903	0.865	0.831	0.741	0.819
	AUROC	0.803	0.802	0.806	**0.809**	0.771	0.677	0.418	0.703
DeepSeek-VL2	Conf	0.572	0.643	0.641	0.651	0.596	0.718	**0.452**	0.896
	Brier	**0.177**	0.182	0.189	0.198	0.224	0.251	0.416	0.372
	ECE	0.053	0.083	0.082	0.097	**0.046**	0.157	0.367	0.341
	AUCPR	**0.862**	0.861	0.852	0.818	0.741	0.708	0.597	0.628
	AUROC	**0.814**	0.813	0.797	0.786	0.679	0.700	0.430	0.599

the correct answer without consulting the image because the question has a canonical or factoid answer. A confidence estimator should not necessarily lower its score on these samples; the answer is correct, just not visually grounded. The clearer test of the grounding-detection mechanism is on the failure subset (image-invariant, incorrect, and confidently predicted), where the model was certain, the model was wrong, and the model was not using the image. Per-LVLM counts for this subset (between 1,121 samples on LLaVA-NeXT-13B and 4,108 on DeepSeek-VL2) are reported in Table 40. Table 43 reports per-LVLM Mean Conf and BS on this subset.

BICR achieves both the lowest mean confidence and the lowest BS on four of five LVLMs (Qwen, InternVL, Gemma, DeepSeek). The LLaVA cell goes to P(True): on this subset (which is incorrect by construction) P(True)'s near-zero confidence happens to coincide with the all-incorrect ground truth, which is the only competitive BS value P(True) attains anywhere in this analysis. On the four LVLMs where BICR wins, its mean confidence on confidently-wrong-and-image-invariant samples is between 0.05 and 0.18 lower than the next-best trainable method, and its BS is between 0.04 and 0.19 lower. The architecturally similar trainable baselines (P(I Know), InternalInspector, SAPLMA) all leave their confidence high on this subset, indicating that the confidence-suppression behavior on the image-invariant failure population is specifically attributable to BICR's rank loss rather than to the shared probe architecture or training data.

J.5 Cross-Subset Win Summary

To check that the headline pattern is not specific to the single $\text{flip}_{\text{swap}} = 0$ cut, we additionally evaluate every method against twelve different ways of identifying image-invariant samples and aggregate the per-cell winners across the full battery. The twelve subsets are: five behavioural-flip variants (the $\text{flip}_{\text{swap}} = 0$ cut itself; an analogous $\text{flip}_{\text{para}} = 0$ cut on the question-paraphrase view; a noise-overlay variant; a multi-proxy consensus combining the swap and paraphrase cuts; and a contaminated control using the blank view, included only as a sanity check on the win-counting procedure);

Table 42: Paired-bootstrap mean and 95% CI of $(\mathrm{BS}^{\mathrm{baseline}} - \mathrm{BS}^{\mathrm{BICR}})$ on the image-invariant subset, $B = 2{,}000$ resamples (RNG seed 23). Positive values mean BICR has lower BS (better calibration) than the baseline on image-invariant samples. **Bold** marks 95% CIs strictly above zero (BICR significantly better calibrated).

Baseline	Qwen3-VL-8B	LLaVA-1.6-13B	InternVL3.5-14B	Gemma3-27B	DeepSeek-VL2
II	**+0.005 [+0.00, +0.01]**	**+0.049 [+0.05, +0.05]**	**+0.014 [+0.01, +0.02]**	-0.007 [-0.01, -0.00]	**+0.005 [+0.00, +0.01]**
PIK	**+0.014 [+0.01, +0.02]**	**+0.022 [+0.02, +0.02]**	**+0.011 [+0.01, +0.01]**	-0.007 [-0.01, -0.00]	**+0.012 [+0.01, +0.01]**
SAPLMA	**+0.030 [+0.03, +0.03]**	**+0.057 [+0.05, +0.06]**	**+0.033 [+0.03, +0.04]**	-0.011 [-0.01, -0.01]	**+0.021 [+0.02, +0.02]**
CCPS	**+0.095 [+0.09, +0.10]**	**+0.064 [+0.06, +0.07]**	**+0.047 [+0.04, +0.05]**	+0.003 [-0.00, +0.01]	**+0.048 [+0.04, +0.05]**
PE	**+0.069 [+0.06, +0.07]**	**+0.056 [+0.05, +0.06]**	**+0.068 [+0.06, +0.07]**	**+0.064 [+0.06, +0.07]**	**+0.074 [+0.07, +0.08]**
P(True)	**+0.469 [+0.46, +0.48]**	**+0.180 [+0.17, +0.19]**	**+0.260 [+0.25, +0.27]**	**+0.401 [+0.39, +0.41]**	**+0.239 [+0.23, +0.25]**
Self-Probing	**+0.094 [+0.09, +0.10]**	**+0.117 [+0.11, +0.12]**	**+0.045 [+0.04, +0.05]**	**+0.079 [+0.07, +0.09]**	**+0.195 [+0.19, +0.20]**

Table 43: Failure subset (image-invariant, incorrect, and confidently predicted): the image-invariant subset further restricted to samples on which the LVLM was both incorrect and assigned the original top-1 token a probability above 0.8. Lower is better for both Mean Conf and BS. Bold marks the best method per (LVLM, metric) cell.

	Qwen3-VL-8B		LLaVA-1.6-13B		InternVL3.5-14B		Gemma3-27B		DeepSeek-VL2	
Method	Conf	Brier	Conf	Brier	Conf	Brier	Conf	Brier	Conf	Brier
BICR	**0.490**	**0.289**	0.709	0.541	**0.484**	**0.292**	**0.442**	**0.275**	**0.471**	**0.291**
II	0.682	0.491	0.770	0.605	0.709	0.519	0.583	0.369	0.547	0.336
PIK	0.673	0.496	0.793	0.661	0.653	0.477	0.579	0.394	0.528	0.336
SAPLMA	0.746	0.600	0.790	0.659	0.719	0.574	0.549	0.357	0.519	0.336
CCPS	0.640	0.445	0.909	0.830	0.770	0.621	0.614	0.422	0.560	0.329
PE	0.869	0.757	0.745	0.559	0.848	0.723	0.873	0.763	0.712	0.511
P(True)	0.517	0.509	**0.364**	**0.151**	0.817	0.790	0.622	0.613	0.483	0.321
Self-Probing	0.911	0.870	0.920	0.865	0.772	0.678	0.873	0.816	0.868	0.799

three continuous variants that take the top-5%, top-10%, and top-25% of samples ranked by the smallest $\mathrm{dp}_{\mathrm{swap}}$ (equivalently, the samples on which image substitution causes the smallest probability drop on the original top-1 token); three confidence- and correctness-conditioned cuts derived from the image-invariant subset (high-confidence, incorrect, and confidently-incorrect); and one auxiliary contaminated cut. Each subset is evaluated on each of the five LVLMs against each of five metrics (Mean Conf, BS, ECE, AUCPR, AUROC), giving $12 \times 5 \times 5 = 300$ possible (subset, LVLM, metric) cells. Twenty cells are dropped because the subset is empty or has fewer than the minimum sample count for the corresponding metric, leaving 280 valid cells.

For each valid cell we identify the method whose value is best on that cell (lowest for Mean Conf, BS, and ECE; highest for AUCPR and AUROC). Table 44 reports the total number of cells each of the eight methods wins, broken down by metric.

BICR is the per-cell winner on 153 of 280 cells (54.6%), more than three times the count of the next-best method. Two of the twelve subsets use the blank-view diagnostic against which BICR is trained, so BICR's wins on those subsets are expected by construction; we retain them in the count for completeness, but excluding them leaves the qualitative ranking unchanged. P(True)'s 41 wins are entirely concentrated in the Mean Conf column (37 of its 41), where its uniformly low confidence is rewarded; on the four metrics that require confidence to be accurate (BS, ECE, AUROC, AUCPR), BICR wins 131 of 240 cells. The win counts disaggregate by LVLM as 30 (Qwen), 33 (LLaVA), 44 (InternVL), 9 (Gemma), and 37 (DeepSeek): Gemma is the weak case, consistent with the headline significance test and with Gemma's behavior in the main results.

J.6 Discussion

The calibration advantage concentrates on the population the rank loss is designed to suppress confidence on, not uniformly across the test set. On the image-invariant subset, BICR's BS is significantly lower than every trained baseline's on 31 of 35 (LVLM, baseline) pairs; on the failure subset (image-invariant and confidently wrong), BICR's mean confidence is 0.05 to 0.18 below the next-best trainable method on the four LVLMs where it wins, with BS gaps of comparable size. The architecturally similar baseline P(I Know) shares BICR's probe family, training data, and search budget, and differs only in the absence of the rank loss against the blank view; on the image-invariant subset it consistently leaves its confidence higher than

Table 44: Per-method per-cell win counts across the cross-subset battery defined in §J.5: 12 subset definitions for image-invariance, 5 metrics (Mean Conf, BS, ECE, AUROC, AUCPR), and 5 LVLMs, for 280 valid cells in total. Each cell contributes one win to the method whose value is best (lowest for Mean Conf, BS, ECE; highest for AUROC, AUCPR). **Bold** marks the most-winning method per metric column and overall.

Method	Mean Conf	BS	ECE	AUROC	AUCPR	Total
BICR	22	**38**	**35**	**28**	**30**	**153**
II	0	4	4	7	6	21
PIK	0	9	5	11	14	39
SAPLMA	0	6	6	3	0	15
CCPS	1	1	6	1	0	9
PE	0	0	2	0	0	2
P(True)	**37**	2	2	0	0	41
Self-Probing	0	0	0	0	0	0

BICR does and is significantly worse-calibrated by BS on four of five LVLMs. InternalInspector and SAPLMA, which use different probe architectures but also do not see the blank-view contrast during training, behave the same way. The pattern is consistent with the rank loss producing the observed effect, and not with a uniform regularisation story.

Image-invariant does not mean wrong: the pooled subset is mostly correct samples, so suppressed mean confidence on it would be miscalibration, not grounding detection. On the pooled subset, $\mathrm{flip}_{\mathrm{swap}} = 0$ samples are on average more accurate than the complement (Qwen 73.5% vs. 65.9%; Gemma 71.1% vs. 59.0%): canonical-answer or factoid questions on which the LVLM produces the right answer without using the image. A well-calibrated method should keep its confidence high on these, and BICR does. The intuition that grounding detection should produce uniformly lower confidence on $\mathrm{flip}_{\mathrm{swap}} = 0$ holds on the failure subset (where the model is wrong and the

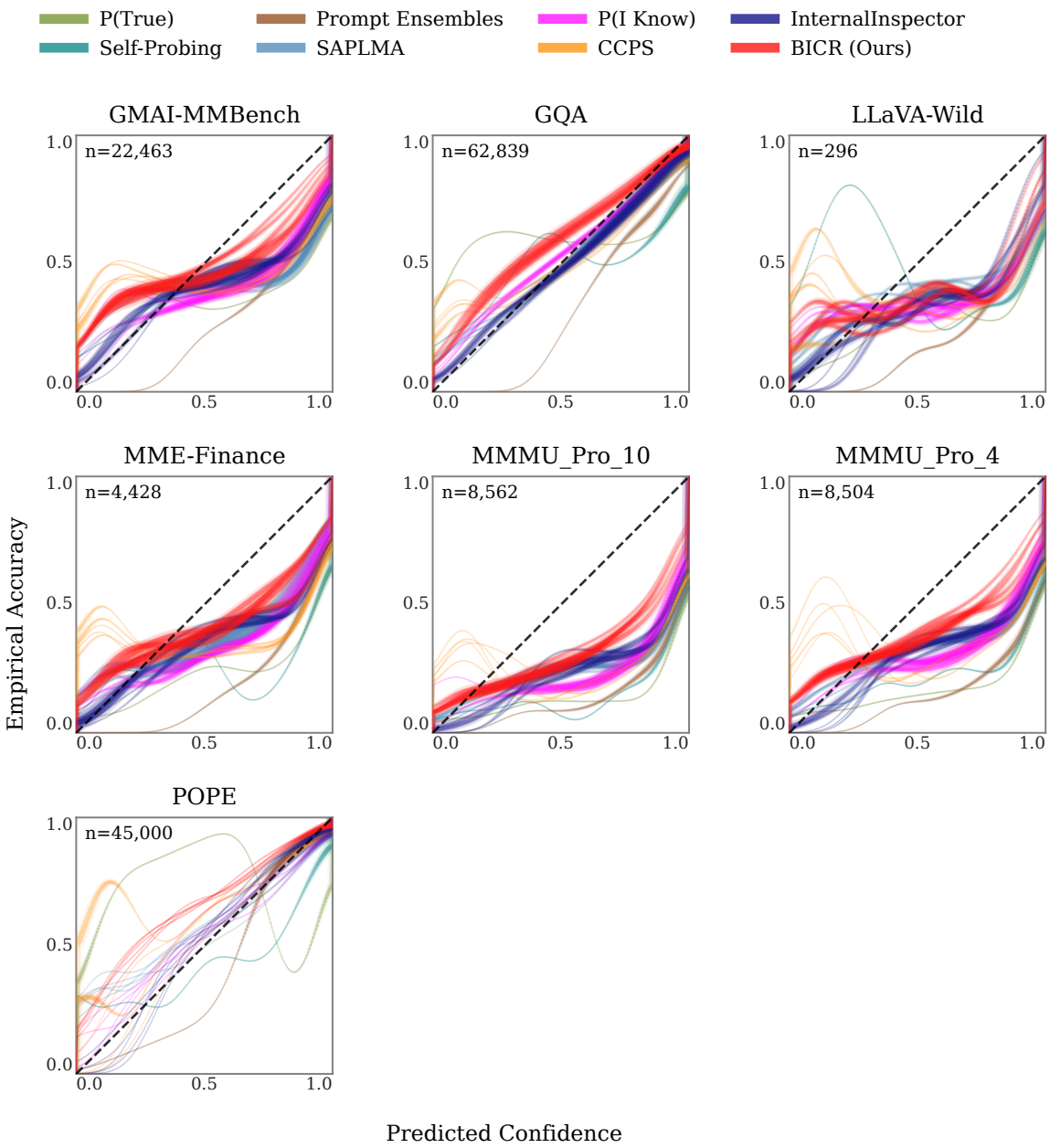


Figure 11: Pooled per-dataset reliability diagrams. Each panel fixes one source dataset and pools every (ground truth, confidence) pair across all five LVLMs (n in panel titles is the total number of pooled test samples). Each method contributes five translucent curves in a method-specific color, one per seed; line thickness within each curve is proportional to local bin density.