Large Vision Language Models as Algorithmic Reasoners for Multimodal Annotations

Shaina Raza[*]
Vector Institute for Artificial Intelligence
Toronto, Canada.
shaina.raza@vectorinstitute.ai

Mahveen Raza[†]
Independent Student Researcher
Toronto, Canada
mahveen.raza10@gmail.com

Abstract

Large vision–language models (LVLMs) can function as algorithmic annotators by not only assigning labels to multimodal inputs but also generating structured reasoning traces that justify those labels. We introduce **Reasoning-as-Annotation (RaA)**, a paradigm in which an LVLM outputs a human-interpretable rationale, calibrated confidence, and evidence pointers alongside each label, effectively acting as both classifier and explainer. We evaluate RaA on bias detection in images using a curated dataset of $\tilde{2}$,000 examples with human gold labels. Across closed- and open-source LVLMs, RaA preserves accuracy relative to black-box labeling while adding transparency: rationales were coherent and grounded in 75–90% of cases, evidence pointers auditable in 70–85%, and confidence scores correlated with correctness ($r=0.60$–0.76). These results show RaA is model-agnostic and maintains predictive quality while producing interpretable, auditable annotations. We position that RaA offers a scalable way to transform opaque labels into reusable reasoning traces for supervision and evaluation.

1 Introduction

Multimodal foundation models integrate information across language, vision, and other modalities, enabling joint perception and reasoning [19]. In text-only domains, prompting methods such as Chain-of-Thought (CoT) [17] and Self-Consistency [16] improve performance by eliciting intermediate steps. In multimodal domains, LVLMs such as LLaVA [12], MiniGPT-4 [20], and Claude 3 Opus [2] extend these abilities to justify answers in visual question answering (VQA), interpret diagrams and charts, and connect visual evidence to textual claims. To date, however, these reasoning abilities have been primarily applied to solving downstream tasks (e.g., QA, captioning, problem solving) [8], rather than to supporting structured annotation and data creation.

We introduce **Reasoning-as-Annotation (RaA)**, a paradigm in which a Large Vision–Language Model (LVLM) acts as a *reasoner-annotator*, as shown in Figure 1. Given an input image, the model (i) inspects salient visual cues, (ii) produces a structured, stepwise rationale grounded in what is depicted, and (iii) outputs a calibrated label with uncertainty, provenance, and pointers to supporting evidence. The central research question is: *Can LVLMs serve as reliable, interpretable, and scalable annotators by making their reasoning explicit?* We argue that explicit reasoning transforms annotation from a black-box decision into an auditable process. In RaA, sampling multiple rationales and aggregating them via self-consistency improves robustness; exposing rationales allows human reviewers to efficiently verify and correct outputs; and storing reasoning traces creates reusable supervision signals for downstream training, such as explanation-tuned models. RaA generalizes to other multimodal reasoning tasks such as VQA, emotion recognition, and content moderation.

[*]
[†]

Multimodal Algorithmic Reasoning Workshop at NeurIPS 2025 (MAR-NeurIPS 2025).

Input V (image)
Cafeteria-like scene: one female-presenting subject is *depicted* sitting at a table; a group of male-presenting subjects stand together and laugh nearby.

LVLM reasoning (RaA)

- Foreground: the seated subject is *depicted* without proximate interlocutors.
- Background: a clustered group appears socially engaged (smiling, facing each other).
- Composition: salience contrast (solo foreground vs. cohesive background) can cue imbalance.
- Interpretation: likely audience takeaway is social exclusion; *alternative explanations (e.g., personal choice, timing) are possible*.

Annotation (RaA output)

Label: LIKELY BIASED
Confidence: $c = 0.81$
Evidence pointers:
R1 (foreground seated subject)
R2 (background clustered group)
Note: assessment targets likely audience framing effect*; it does not attribute intent or preference to depicted individuals.*

Figure 1: **RaA pipeline.** Inputs (image/text) → LVLM. Black-box: label only. RaA: *rationale*, *label* (Likely/Unlikely Biased), *confidence* (self-consistency), and *evidence pointers*. Right: evaluation metrics (Accuracy, Rationale Quality, Confidence Calibration, Auditability).

RaA builds directly on two strands of reasoning research: CoT prompting [17] and Self-Consistency [16], which have proven effective in text-only LLMs, and multimodal rationalization techniques developed in LVLMs such as LLaVA and MiniGPT-4. We extend these reasoning mechanisms from task-solving to annotation, by requiring models to output structured rationales, calibrated confidence, and auditable evidence pointers alongside labels. In this work, we demonstrate RaA on curated and synthetic visual examples that are designed to stress-test how models reason about compositional cues. These examples include everyday scenes where framing can suggest imbalance (e.g., a single individual foregrounded against a cohesive group), even when no explicit textual content is present. While these demonstrations validate the paradigm under controlled conditions, the framework is task-agnostic and can be applied to domains such as safety analysis in VQA, multimodal sentiment classification, or medical image triage. We position LVLMs as reasoner-annotators that enable transparent, auditable, and scalable annotation.

2 Related Work

Reasoning in Language Models. Prompting methods CoT [17] and Self-Consistency [16] have demonstrated that eliciting intermediate reasoning steps improves robustness and accuracy in text-only LLMs. Follow-up work has explored program-aided reasoning [10], tree-of-thought search [18], and self-reflection strategies [14], all of which emphasize interpretability through explicit reasoning traces. RaA builds directly on this line of research by repurposing CoT and Self-Consistency for annotation rather than task solving. **Multimodal LVLM Reasoning.** Recent LVLMs extend reasoning into the visual domain, where models justify predictions with textual explanations. LLaVA [12], MiniGPT-4 [20], Claude 3 Opus [2], GPT-4V, and Gemini 2.0 [7] demonstrate that multimodal reasoning improves performance in VQA, chart interpretation, and image captioning. However, these systems use reasoning primarily for inference. In contrast, RaA reframes multimodal reasoning as an annotation protocol, structuring outputs into rationales, confidence estimates, and evidence pointers. **Annotation Frameworks.** Conventional annotation pipelines often rely on human-in-the-loop methods, active learning [11], or crowd-sourcing [5]. While effective for scaling data, these pipelines typically produce opaque categorical labels without interpretability. More recent work on explanation-based annotation [13] proposes aligning labels with model explanations, but has largely focused on text. RaA differs by explicitly positioning LVLMs as *reasoner-annotators* that generate transparent, auditable annotations in multimodal domains. So, prior work has either (i) improved model performance through reasoning, (ii) applied reasoning to multimodal inference tasks, or (iii) designed human-centered annotation pipelines, RaA is, to our knowledge, the first systematic attempt to unify these threads into a single framework.

3 Method

Reasoning-as-Annotation (RaA) Overview We formalize **RaA** as a pipeline in which a LVLM acts as a *reasoner-annotator*, as shown in Figure 2. For an input x (image-only here, but it is extensible to

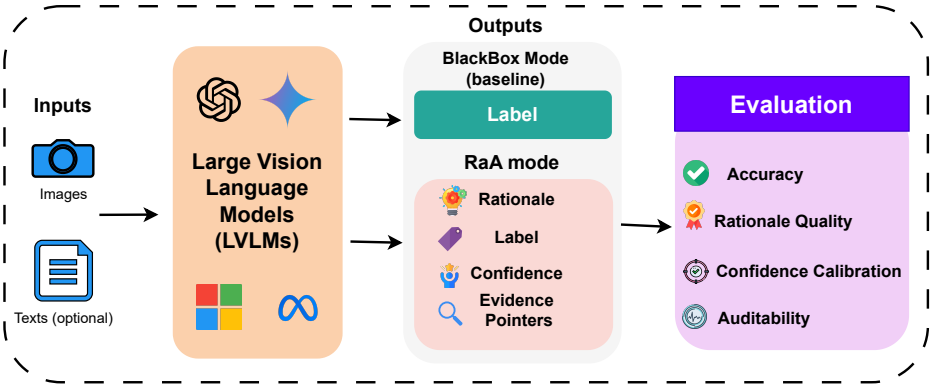

Figure 2: **RaA pipeline and evaluation.** Inputs → LVLM → (Label-only baseline) or (RaA: Rationale, Label, Confidence, Evidence) → Evaluation on Accuracy, Rationale Quality, Confidence Calibration, and Auditability.

(T, V)), the model first generates an explicit rationale R and then emits an annotation $A = (y, c, E)$ consisting of a label y, a confidence score $c \in [0, 1]$, and evidence pointers E. By externalizing R and E, RaA transforms opaque decisions into auditable annotations.

Rationale Generation & Prompting Before producing A, the LVLM is instructed to generate a structured rationale R following a fixed schema: (i) foreground description, (ii) background description, (iii) composition/interaction, and (iv) brief interpretation. This mirrors CoT style intermediate steps while grounding explanations in visual evidence. We show this schema via prompt templates provided in Appendix A.

Confidence Estimation RaA assigns a confidence c using a self-consistency procedure. For each input, we sample multiple reasoning paths under mild decoding stochasticity; each path yields (R, y). Let k be the number of samples and let $\hat{y}$ be the majority label. We define $c = \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}[y_i = \hat{y}], \quad \hat{y} = \text{mode}(y_{1:k})$ so c measures agreement across diverse reasoning paths. Hyperparameter choices (e.g., k) are reported in Section D.1. Alongside R and y, the model outputs coarse evidence pointers E that link the rationale to salient visual regions (e.g., "foreground subject," "background group"). The goal is auditability, not dense localization. Each annotation is stored as (y, c, R, E) keyed by input identifier. The formal schema is compact by design to enable reuse for explanation-tuned training and human auditing; a JSON example is provided in Appendix A.3.

4 Experiments and Results

Experimental Setup **Data Source and Annotations.** We evaluate RaA on an image-only collection of approximately $N \approx 2000$ examples, constructed from two sources: (i) a curated subset from prior internal annotation studies and (ii) synthetic images designed to stress-test compositional cues (e.g., a foregrounded individual contrasted with a cohesive group). This dataset is used as a methodological testbed rather than a benchmark. For evaluation, we obtained human gold labels on the same bias-detection task (`Likely Biased` vs. `Unlikely Biased`), as discussed in Appendix C. Human annotations were not used to prime or guide model outputs; they serve solely as reference labels for accuracy and as the basis for auditing rationales and evidence pointers.

Models and Evaluation Protocol We evaluate the RaA framework across a diverse set of LVLMs spanning both proprietary and open-source ecosystems. We evaluate closed-source systems GPT-4o [9] and Gemini 2.0 Flash [7], as well as open-source models including Phi-4 [1], LLaMA-3.2 11B Vision-Instruct [3], Qwen2.5-7B [4], and InternVL2.5 [6]. Each model was evaluated under two conditions: (i) *black-box mode*, where the model outputs a direct label, and (ii) *RaA mode*, where the model is prompted to output a rationale, label, confidence score, and evidence pointers. For RaA, we sample $k = 5$ reasoning paths and compute confidence c as the fraction of consistent labels. We use four metrics: two human-judged: **Rationale Quality** and **Auditability** (n=200, two raters; we report Cohen's κ), and two automatic: **Accuracy** (vs. gold labels) and **Confidence Calibration** (correlation between c and correctness, with reliability plots/Brier score in Appendix D). Together, these capture predictive performance and interpretability.

Table 1: Comparison of LVLMs under *Black-box* (label-only) and *RaA* (reasoner-annotator) modes. Accuracy is measured against human gold labels. Rationale quality, confidence–correctness correlation, and auditability are reported only for RaA. Models are grouped into closed and open-source.

Model	Accuracy (%)		Rationale Quality (%)	Conf.–Correct Corr.	Auditability (%)
	Black-box	RaA			
			Closed-source		
GPT-4o [9]	91	92	90	0.76	85
Gemini 2.0 Flash [7]	89	90	88	0.73	82
Aggregate (closed-source)	*90*	*91*	*89*	*0.75*	*84*
			Open-source		
Phi-4 [1]	85	86	82	0.68	77
LLaMA-3.2 11B [3]	84	85	81	0.66	75
Qwen2.5-7B [15]	80	81	76	0.60	70
InternVL2.5 [6]	83	84	79	0.65	73
Aggregate (open-source)	*83*	*84*	*80*	*0.65*	*74*

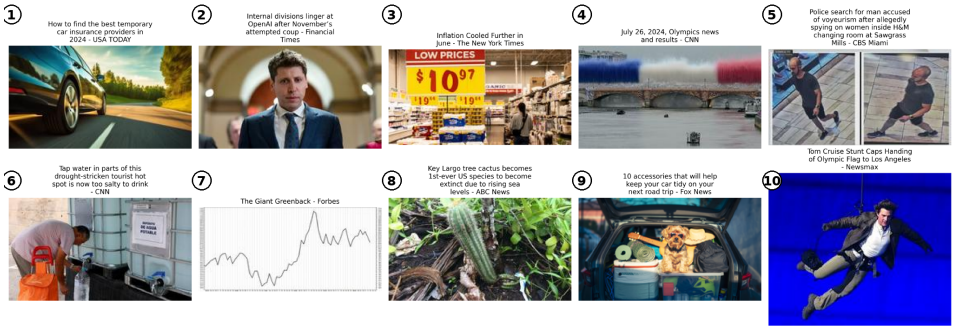

Figure 3: **Qualitative RaA on curated images.** Numbers index tiles; see Appendix Table 6 for selected cases. We avoid identifying individuals and use small, transformative crops for research use.

Analysis Table 1 shows that RaA maintains accuracy relative to black-box predictions while adding structured reasoning, confidence, and evidence that make annotations interpretable and auditable. Accuracy is broadly preserved across all models, with closed-source systems achieving the highest scores. In addition, RaA produces high-quality rationales, calibrated confidence estimates, and verifiable evidence pointers, none of which are available in black-box mode. Three main findings emerge: (1) **Accuracy is preserved.** Across all evaluated models, requiring reasoning steps does not degrade predictive quality. (2) **Interpretability is improved.** Rationales were coherent and grounded in 75–90% of RaA outputs, depending on model. Evidence pointers were auditable in 70–85% of cases, enabling rapid human validation. (3) **Confidence is informative.** RaA confidence scores correlated strongly with prediction correctness ($r = 0.60$–0.76); low-confidence cases ($c < 0.5$) often corresponded to ambiguous or borderline labels, guiding reviewers to uncertain items.

5 Conclusion and Limitations

We presented RaA, a framework that positions LVLMs as reasoner-annotators capable of producing not only labels but also structured rationales, calibrated confidence scores, and evidence pointers. Our experiments across closed- and open-source LVLMs show that RaA preserves predictive accuracy while substantially improving interpretability and auditability. By externalizing reasoning, RaA turns annotation into a transparent, accountable process, enabling human reviewers to verify outputs more efficiently and providing reusable supervision signals for downstream model training.

Despite these promising results, our study has several limitations. First, the evaluation dataset was limited to $\sim$2,000 curated and synthetic image-only examples, which may not capture the full diversity of real-world multimodal bias. Second, RaA relies on prompt engineering and structured templates, making outputs sensitive to design choices and decoding parameters. Third, while rationales were often plausible, they occasionally reflected superficial or spurious cues, raising questions about faithfulness. Finally, we focused on short-form bias annotation; broader applications such as medical or safety-critical domains require careful validation before deployment. To facilitate reuse while

respecting licensing, we plan to release (i) prompts and (iii a 200-image subset with permissive licenses. Future work should extend RaA to larger and more diverse multimodal datasets.

References

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Anthropic. Claude: a family of large language models, 2025. Accessed: 2025-08-22.

[3] A. at Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Technical report, Meta, 2024.

[4] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] K. Benoit, D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov. Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data. In *American Political Science Review*, volume 110, pages 278–295, 2016.

[6] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[7] G. Cloud. Gemini 2.0 Flash, Apr. 2025. Generative AI on Vertex AI documentation. Last updated 2025-04-23.

[8] M. A. Ferrag, N. Tihanyi, and M. Debbah. Reasoning Beyond Limits: Advances and Open Problems for LLMs, Mar. 2025. arXiv:2503.22732 [cs].

[9] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[10] A. Kabra, S. Rangreji, Y. Mathur, A. Madaan, E. Liu, and G. Neubig. Program-aided reasoners (better) know what they know, 2023.

[11] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, and M. Okumura. A survey on deep active learning: Recent advances and new frontiers, 2024.

[12] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[13] S. Shimizu, H. Shohei, Y. Uno, S. Yada, S. Wakamiya, and E. Aramaki. Exploring LLM annotation for adaptation of clinical information extraction models under data-sharing restrictions. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14678–14694, Vienna, Austria, July 2025. Association for Computational Linguistics.

[14] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[15] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[16] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[18] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[19] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

[20] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Appendix

Broader Impact. RaA highlights both opportunities and risks in using LVLMs for annotation. On one hand, explicit rationales can enhance transparency, facilitate auditing, and reduce reliance on costly manual labeling. On the other hand, generated rationales may give a false impression of reliability if users conflate plausibility with truthfulness. Deploying RaA in sensitive domains requires careful oversight, robust human-in-the-loop practices, and safeguards against reinforcing existing biases. By foregrounding reasoning in annotation, RaA encourages more accountable use of LVLMs, but its broader societal value depends on ensuring that explanations remain faithful, fair, and responsibly applied.

A Prompt Templates

We include here the exact prompts used to instantiate Reasoning-as-Annotation (RaA). Prompts were designed to enforce a structured reasoning schema, requiring the LVLM to produce rationales prior to labels.

A.1 Image-only RaA Prompt

```
SYSTEM:
You are an annotator for Reasoning-as-Annotation (RaA). Follow instructions
    exactly. Be concise, neutral, and avoid attributing intent. Base all
    statements only on what is depicted. Return ONLY a single JSON object that
    conforms to the required schema. Do not include any extra commentary.

USER:
Task: Given an image, (1) describe the foreground elements, (2) describe the
    background elements, (3) explain how foreground and background interact
    compositionally, and (4) provide a short interpretation of what this
    composition may suggest for likely audience framing.

Then output a JSON object with the following fields:
- "label": one of ["Likely Biased","Unlikely Biased"]
- "confidence": a float in [0,1] reflecting your certainty (e.g., 0.81)
- "rationale": a single string concatenating four parts in this order:
   "Foreground: ... Background: ... Composition: ... Interpretation: ..."
- "evidence": a list of short pointers to salient regions (e.g., ["foreground
    subject","background group"])

Constraints:
- The JSON must be valid and parseable.
- Use only the two label values above.
- Do not add fields beyond the schema.
- Do not include markdown or explanations outside JSON.

Now perform the analysis for the image and return ONLY the JSON object.
```

A.2 Image+Text RaA Prompt

```
SYSTEM:
You are an annotator for Reasoning-as-Annotation (RaA). Follow instructions
    exactly. Be concise, neutral, and avoid attributing intent. Base visual claims
    on the image and textual claims on the provided text. Return ONLY a single
    JSON object that conforms to the required schema. Do not include any extra
    commentary.

USER:
Task: Given (Text, Image), (1) identify the main claim or framing in the text, (2)
    describe salient foreground and background elements in the image, (3) explain
    how the text and image interact (e.g., reinforce, contradict, frame), and (4)
    provide a short interpretation of what this multimodal pairing may suggest for
    likely audience framing.

Then output a JSON object with the following fields:
- "label": one of ["Likely Biased","Unlikely Biased"]
- "confidence": a float in [0,1] reflecting your certainty (e.g., 0.81)
- "rationale": a single string concatenating five parts in this order:
   "Text: ... Foreground: ... Background: ... Interaction: ... Interpretation: ..."
- "evidence": a list of short pointers (e.g., ["text span: <brief
    quote>","foreground figure","background group"])

Constraints:
- The JSON must be valid and parseable.
- Use only the two label values above.
- Keep quotes in evidence short (avoid long excerpts).
- Do not add fields beyond the schema.
- Do not include markdown or explanations outside JSON.

Now perform the analysis for the given text and image and return ONLY the JSON
    object.
```

A.3 Schema

```
{
  "input_id": "img_00123",
  "label": "Likely Biased",
  "confidence": 0.81,
  "rationale": "Foreground subject is seated without proximate interlocutors; \
background group appears cohesive and socially engaged; composition emphasizes \
contrast between isolation and group cohesion; likely audience framing suggests
imbalance.",
  "evidence": ["foreground subject", "background group"]
}
```

Table 2: RaA output schema (per input).

Field	Type	Description
`input_id`	string	Unique identifier of the input image/example.
`label` (y)	enum	`Likely Biased` / `Unlikely Biased`.
`confidence` (c)	float	Self-consistency agreement in $[0, 1]$.
`rationale` (R)	string	Structured explanation (FG/BG/Composition/Interpretation).
`evidence` (E)	list[string]	Coarse pointers to salient regions (e.g., "foreground subject").

B Dataset Statistics

We report distributions of the image-only collection used in RaA experiments. The dataset totals $N \approx 2000$ examples, comprising both curated and synthetic sources. Gold labels were obtained from human annotators for evaluation only.

Table 3: Distribution of RaA dataset (image-only).

Source	Count	Likely Biased (%)	Unlikely Biased (%)
Curated (internal)	1200	54	46
Synthetic (stress-test)	800	61	39
Total	2000	57	43

C Annotation Protocol

Three annotators with graduate-level training in computer science and media studies were recruited from within our research group. Annotators were instructed to read the task description carefully and classify each example as `Likely Biased` or `Unlikely Biased`. The instructions emphasized that (i) annotators should base judgments strictly on the depicted composition (foreground/background contrast, framing cues, omission of context), (ii) they should avoid attributing intent or psychological states to individuals in the images, and (iii) when uncertain, they should select the label that best reflects the likely audience-facing framing effect.

Each example was labeled independently by all annotators. Disagreements were resolved by majority vote, with ties adjudicated in discussion. Inter-annotator agreement across the full set ($N = 2000$) was substantial ($\kappa = 0.78$). These gold labels are used exclusively as reference for accuracy evaluation and auditing of model rationales and evidence pointers.

D Evaluation Protocol

Rationale Quality (RQ). On a stratified sample of $n = \mathbf{200}$ model outputs, two trained raters judged whether the RaA rationale is (i) coherent, (ii) grounded in the input, and (iii) free of hallucinated entities. A rationale counts as "high-quality" if it satisfies all three criteria. We report the proportion and inter-rater agreement (Cohen's κ). Disagreements were resolved by a third rater.

Auditability (AU). For the same $n = \mathbf{200}$ items, raters judged whether evidence pointers E align with the salient regions/spans referenced in the rationale R. We compute AU as the proportion of items marked "aligned." We report κ for the binary aligned/not-aligned decision.

We detail the sample sizes and protocols used for each evaluation metric.

Table 4: Evaluation metrics, reference sources, and sample sizes.

Metric	Reference Source	Sample Size	Description
Accuracy	Human gold labels	$N \approx 2000$	Proportion of predicted labels matching ground truth.
Confidence calibration	Human gold labels	$N \approx 2000$	Pearson correlation between RaA confidence and correctness.
Rationale quality	Human spot-checks	$n = 200$	Judged coherent and grounded across FG/BG/Composition/Interpretation.
Auditability	Human spot-checks	$n = 200$	Evidence pointers aligned with salient input regions.

Rationale quality and auditability were judged by annotators on a random subset of 200 examples. Each case was independently reviewed by two annotators, with disagreements resolved by majority vote. Accuracy and calibration were computed on the full gold-labeled set ($N \approx 2000$).

D.1 Hyperparameter Settings

We detail the hyperparameters used in RaA prompting and inference. Unless otherwise noted, values were tuned empirically for stability rather than performance optimization.

Table 5: Hyperparameters used in RaA experiments.

Parameter	**Value**	**Description / Rationale**
Self-consistency samples k	5	Each input decoded k times with mild stochasticity; confidence c is the fraction of samples agreeing with the modal label.
Decoding temperature	0.7	Balances diversity and faithfulness in rationale generation.
Top-p	0.9	Encourages variability in reasoning paths while limiting tail tokens.
Max tokens (per rationale)	256	Prevents overlong rationales while allowing the full schema.
Stopping criterion	Schema delimiter	Decoding halts when the structured schema is completed.
Batch size	8 per GPU (open-source); N/A (APIs)	Open models run batched; hosted APIs are evaluated sequentially per provider constraints.
Precision	BF16 (open-source); N/A (hosted APIs)	Mixed precision for efficiency on open models; API precision is not user-configurable.
Prompt schema	Fixed template	*Image-only:* FG $\rightarrow$ BG $\rightarrow$ Composition $\rightarrow$ Interpretation. *Image+text:* Text $\rightarrow$ FG $\rightarrow$ BG $\rightarrow$ Interaction $\rightarrow$ Interpretation.

E Qualitative Examples

Figure 3 illustrates how RaA adds actionable signal beyond label-only outputs. In tile 3 (grocery price placard) RaA predicts *Unlikely* with moderate confidence (c=0.64), reflecting neutral retail imagery. Tile 5 (potable-water station) is *Likely* (c=0.78) due to problem-forward framing around scarcity. Tile 6 (economics chart) is *Unlikely* (c=0.60), consistent with descriptive analysis. Tile 8 (endangered cactus close-up) is *Likely* (c=0.74) given evocative conservation framing, while tile 9 (car-tips packing shot) is *Unlikely* (c=0.69) with service-oriented tone. Across examples, low c aligns with borderline cases and naturally surfaces items for human review; evidence pointers (tags/spans/regions) make these judgments verifiable without revealing identities.

Table 6: Selected cases from Fig. 3.

#	RaA Label (c)	Evidence	One-line rationale
3	Unlikely (0.64)	price placard, aisle shelf	Neutral headline; generic low-price tag, no partisan cue.
5	Likely (0.78)	water sign, container tap	"agua potable" signage; scarcity/problem-forward framing.
6	Unlikely (0.60)	trend plot axes	Descriptive economic chart; minimal emotive language.
8	Likely (0.74)	cactus foreground	Conservation framing + evocative close-up.
9	Unlikely (0.69)	trunk storage items	Service-oriented tips; practical tone.