International Journal of Computers and Informatics (IJCI)
Vol. (5), No. (11)
IJCI
November 2026
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (5)، العدد (11)

Evidence-Constrained Visual Question Answering: A Diagnostic Study of Grounding and Selective Prediction

Hamsa M. Ahmed
Department of Islamic Finance and Banking Sciences, College of Islamic Sciences,
University of Anbar, Iraq
hamsa.m.ahmed@uoanbar.edu.iq

Shokhan M. Al-Barzinji
Department Theory, College of Physical Education and Sports Sciences,
University of Anbar, Ramadi, Iraq
shokhan.albarzinji@uoanbar.edu.iq

Abstract

Inspectable visual evidence does not necessarily indicate that an evidence-based acceptance rule is more reliable in prediction. The aim of this study is to measure that difference in binary object-presence visual question answering. It is accompanied with an independent object detector providing localized evidence for positive answers, and an open multimodal generator, SmolVLM-256M-Instruct. The number of positive candidates accepted depends on a development-selected threshold, while negative candidates abstain because lack of detection cannot establish absence. The exploratory pilot resulted in a new replication with 60 new COCO val2017 images and 360 new adversarial POPE questions, where the images were split into development/evaluation sets. 68.75% reference-label agreement on 240 evaluation questions was obtained using direct generation. The evidence gate accepted 39 answers at 16.25% coverage with 100.00% reference agreement, while positive-only confidence selection achieved 97.44% agreement at the same accepted count. The gate had a higher accepted-answer accuracy point estimate than positive-only confidence selection, but the exploratory difference interval included zero; superiority was not established. Also reported in this study are threshold sensitivity, positive-answer retention, image-cluster bootstrap intervals and checks on component training provenance. The results apply to just one backbone and a limited object-presence task, and do not establish semantic evidence sufficiency, causal visual dependence or general hallucination mitigation. The contribution is a reproducible diagnostic case study that demonstrates how answer-class composition and component training provenance should be controlled when assessing the evidence-constrained visual answering problem.

Keywords: Visual Question Answering, Multimodal Artificial Intelligence, Visual Grounding, Object Hallucination, Selective Prediction, Evidence Verification, Reproducibility.

1. Introduction

Multimodal Artificial Intelligence is the use of computational tools that integrate the visual and linguistic aspects of a picture and can respond to questions about the image content. The prediction of answer from an image and a natural language question is termed as visual question answering (VQA) [1] formalizes this interaction. Relevant observations include objects, attributes, quantities, relationships, or readable text. Use examples range from support for blind people asking questions about common photos [3]. In such cases, the answer ought to be in keeping with the image on offer, not just look like a feasible answer.

Reliability demands that the decision to answer be evaluated as well as the information that backs it. A model may generate an incorrect response with confidence and/or a correct answer by means of linguistic regularities. The complementary-image evaluation in VQA v2 emphasizes the need for answers to depend on the relevant image content [2] and the object-hallucination evaluation focuses on the evaluation of statements that include objects that are inconsistent with the reference image [6]. An evidence record gives the inspection a basis, but a linked region may be irrelevant, or not be understood. Improved reliability must be demonstrated through controlled comparisons rather than inferred from an evidence record.

In this study, the VQA questions are grouped into five categories of operations: object presence, attributes, counting, spatial or relational reasoning, and text reading. The categories can overlap and could be different types of evidence required. Presence requires establishing a visible entity, the attribute is about attaching a property to the intended entity, the count is about having distinct instances, the relations are about jointly identifying the relevant entities, and the reading of texts requires recognizable characters. The focus of the present experiments is limited to binary object-presence questions. A positive detection allows local support for the presence of the object, while failure to detect an object cannot establish its absence in cases of occlusion, limited resolution or imperfect detector recall.

This work is related to benchmark VQA, transferable vision-language learning and selective prediction. Image understanding [1, 2] set the task and its significance at the beginning of VQA and bias-reduction benchmarks. Transferable image-language capabilities were expanded with the CLIP and visual instruction tuning [5, 7]. POPE and Hallusion Bench implemented complementary hallucination diagnostics [6, 8]. There are many works that explicitly explore the decision to withhold a response (abstain) [15] and others that investigate how to obtain confidence in a decision and collect evidence for selective visual reasoning [16, 17]. These developments suggest considering the evaluation of an evidence gate under selection based on confidence, when the coverage and the answer classes are similar.

We describe an empirical study of a lightweight evidence-constrained answering procedure, which is diagnostic, in this paper. We combine an open multimodal generator with an independent detector while keeping the detector's bounding-box evidence, and compare selective outputs with controls

ranked by the generator's first-token confidence. An exploratory pilot is followed by another image-disjoint replication, limited to the detector's published validation-image pool. The study questions whether the gate enhances the selection of the positive answers, whether it eliminates reference-disagreeing positive answers, and whether it simply discards correct positive answers, and how answers differ from various acceptance thresholds. We report on observed outcomes without the need for improvement, and differentiate evidence records that are repeatable from those that have been verified as having a semantic grounding.

2. Historical Background and Related Work

2.1. VQA Benchmarks and Visual Dependence:

Antol et al. structured VQA based on open-ended questions and images [1]. Goyal et al. later focussed on image understanding with complementary images with different answers to the same question [2]. This step is important because the extent of agreement in responses does not provide an indication of whether or not the appropriate visual information was utilized. GQA also introduces the concept of compositional reasoning [4] in the evaluation, while VizWiz introduces real-world answerability challenges [3]. The benchmarks set forth are different requirements and mastery of the binary presence judgments does not mean ability to count, relate, or recognize inaccessible image details.

2.2. Multimodal Learning and Hallucination Diagnosis:

Given a pair of visual and linguistic data, CLIP learns transferable representations [5]. Visual Instruction Tuning is a technique that embeds image encoders into language models for interacting with visual data [7]. However, with such powers the threat of object hallucination is still a real problem in systems evaluated, and this is why POPE's polling-based presence questions, [6] were devised. A more general theme of visual illusion and language hallucination is explored by Hallusion Bench [8]. All questions on our experiments are adversarial POPE presence questions, and so the results must not be assumed to apply to the rest of the failure modes included in these benchmarks.

There are several approaches to mitigation that can interfere with visual processing and attention such as Clear Sight and attention-space guidance [9, 10]. This study evaluates an external acceptance policy instead of claiming an improvement to model representations or decoding. We are not proposing to build those mitigation systems, nor claiming to be better than them – the comparison is made at the level of a single generator and the question is around the added -value of detector evidence.

2.3. Selective Answering and Evidence Collection:

Whitehead et al. provide a reliable VQA based on risk and coverage, so that if the answer is unreliable, it is possible to abstain [15]. Eisenschlos et al. look at confidence and calibration of multimodal selective answer [16]. This indicates that it is important to have confidence-based controls: an external gate should not be given credit for reliability that simpler selection could get.

ReCoVERR gathers visual-related evidence to provide fewer unnecessary abstentions [17]. EVICT uses counterfactual masking of regions cited as evidence to assess visual evidence sensitivity [18]. Therefore, evidence acquisition, selective VQA, and evidence-sufficiency probing are done before the present study. We do not claim to be the inventors of the evidence-based answering, or of a new general-purpose architecture, our contribution is a limited diagnostic case study. We isolate answer-class composition, report positive retention/rejection and audit the training provenance of external detector. This focus is on the interpretation of a lightweight gate under explicit constrained conditions. Our detector audit does not measure causal dependence on cited pixels.

3. Research Questions and Scope

RQ1: is there an improvement in accepted-answer reference agreement if using detector-gated positive answering instead of positive-only confidence selection at the same accepted count?

RQ2 asks whether the gate removes positive answers that disagree with the reference or mainly decreases the coverage of correct positive answers.

RQ3 investigates how the observed trade-off evolves for a range of prespecified detector thresholds and when the detector-image provenance is controlled.

Empirical unit: binary object-present question with an image. A reference answer is yes or no. Reference agreement measures agreement with POPE labels, not independent judgment of whether a particular bounding box supports an assertion. The paper compares the performance of a fixed pretrained generator and detector, without training them. It does not support general atomic-claim decomposition, three-way semantic entailment, text recognition or relational verification.

We take “verifiable” to mean that a reader can inspect the source image, question, candidate answer, detector class, confidence score, bounding box and acceptance decision. This operational traceability is less than a proof of correctness. It is not sufficient to take a candidate out of a process to make a positive gate a reliable absence verifier, and a wrong, or irrelevant, box can remain in an auditable record.

4. Methodology

4.1. Generator and Observation Extractor:

The generator used is Hugging FaceTB/SmolVLM-256M-Instruct [11]. Images and original questions are provided by its chat template with the fixed instruction ‘Answer only yes or no'. Inference runs using just two PyTorch threads, greedy decoding, and up to eight generated tokens per inference on the CPU, using float 32 precision. Images are processed with longest edge of the image set to 512 pixels and image splitting disabled. The output parser accepts an initial "yes" or "no" word and treats outputs without either initial word as operational abstentions. The generator isn't given a reference label.

The evidence extractor used is torchvision Faster R-CNN MobileNet V3 Large 320 FPN with its published COCO weights [12]. It is applied once for every image, and returns labels, confidence scores, and boxes with the original image coordinates. All of the classes in the COCO class vocabulary are compared with the noun being queried using the published alias map. Matching boxes are scored the maximum of the matching boxes supplied the evidence score for the gate; if there is no match, score is zero. The detector score is not a probability that the final answer is true – it is merely a ranking signal that is not calibrated.

4.2. Acceptance Procedure:

The generator generates a candidate for each question. If it is positive, the system checks whether a box matching the queried object class reaches the development-selected threshold. A qualifying positive is accepted and associated with the matching boxes. Negative candidates and unmatched positive candidates abstain. This is a scope restriction, not a claim that negative answers are generally unreliable.

Let a_i be the candidate that is parsed for question i and e_i be the maximum matching detection score. Accepting is:

$$g_i = 1(a_i = \text{yes}) \times 1(e_i \geq \tau)$$

Here 1(·) is an indicator and tau is the development-selected threshold. No matching box gives e_i = 0. The policy accepts neither unparsed outputs nor negative answers.

The implementation stores the source row, question, image identifier and hash, raw output, answer (parsed), the confidence in the first token, matching boxes, maximum matching score, split assignment, and processing time. Normalized confidence of the first token is used only for ranking baseline selections; it is the first-token confidence. It is not calibrated and does not imply total correctness of the answers.

Algorithm 1. Executed Evidence-Gated Inference:

1. Place the source image and question in the model input; not include the reference answer.
2. Execute the independent detector once for this image, keeping the boxes and labels.
3. Make a guess with the help of a question prompt and decoding rules.
4. Parse the candidate and resolve the queried object to the detector vocabulary.
5. Only accept a positive candidate who has a matching box that is at or above the frozen threshold.
6. Otherwise abstain; save the decision and evidence record for audit.

4.3. Threshold Selection and Controls:

The candidate thresholds are 0.3, 0.5, 0.7, and 0.9. For development images, select the threshold that would be most likely to allow the maximum number of answers while maintaining observed acceptable precision of at least 0.9. Equal accepted counts count towards higher threshold. If none of

the thresholds are met, record 0.9; and report that the development objective was not met. It is an empirical development rule and not a finite sample guarantee of 90% precision in the population. No selection and threshold revision is made via evaluation labels.

Direct answering is that all candidates generated are evaluated. The all-answer confidence control selects the highest-confidence parsed candidates at the gate's evaluation accepted count. Positive-only confidence control chooses the same number from the positive candidates only. Ties - Stable source-row order. The positive-only control is offered to compensate for the fact that a positive gate alters the composition of the answers. Matching counts helps achieve equal coverage, but does not ensure that the same questions or equally challenging subsets are chosen.

The complete direct+ pool is also reported. The effectiveness of the gate for suppressing the false positive or simply eliminating the correct positive response can be determined by comparing the number of positives before and after the gate. In case of a confidence control, no additional model calls are necessary, but the gate has to be called and costs detector.

5. Experimental Design

5.1. Exploratory Pilot and Provenance Audit:

The first 180 rows of the public POPE mirror [13] were used for the original pilot, for 30 images. A split with 10 image-level development images and 20 image-level evaluation images was used and seeded. This was a convenience sample that was only a test of the pipeline and was not a benchmark for generalisation. The official COCO val2017 archive directory was subsequently audited by identifier. Of these 20 evaluation images, just 5 were in val2017. COCO's 2017 split reorganization makes a val2014 filename insufficient to establish exclusion from the 2017 training split [14]. The pilot is therefore kept as exploratory evidence but isn't used as a detector-unseen evaluation.

5.2. Prespecified Validation-Split Replication:

The replication protocol was drafted after the pilot and before new replication predictions. All 3000 rows of sources were scanned. The images were required to be part of the official COCO val2017 image identifier list (which excludes the 30 images in the pilot list). The number of images that could be used was 66. They were sorted, then shuffled with randomness seed 20261112 and the first 60 were picked. The first 20 images were used for development, and the next 40 for evaluation. All six existing adversarial questions per image were retained, yielding 120 development questions and 240 evaluation questions. The source cache, selection record and the image hashes remain.

The detector's referenced training split is not included in this design, but it is also exposed to detector model selection in its validation split as published in the COCO partition. It also doesn't assert that the generator did not see these images or annotations at all in the pre-training process. Therefore, we do not refer to it as a sequestered unseen-model test, but rather as a validation-split replication. New development data are used to establish a new threshold; the generator, detector, parsing, prompt and threshold candidates are unchanged from the pilot.

5.3. Outcomes and Uncertainty:

Coverage is the number accepted divided by the total number N of questions evaluated. Let l_i be the binary reference loss of question i. Coverage and selective risk are:

$$\text{Coverage} = \frac{1}{N}\sum_i g_i$$

$$\text{Risk} = \frac{\sum_i g_i \ell_i}{\sum_i g_i}, \sum_i g_i > 0$$

Accepted-answer accuracy, reference agreement is the agreement between accepted answers, which equals to 1 minus binary selective risk. If no answers are accepted, accepted-answer accuracy and selective risk are undefined. The fact that the gate is only used for positive answers means that the accepted accuracy of this gate is positive predictive value, which is different from balanced binary classification.

The primary effect is the difference in accepted-answer accuracy between the gate and positive-only confidence selection at the same count. A paired image-cluster bootstrap resamples evaluation images sampling with replacement and all the questions are included. The development threshold is set, the baseline is calculated (recalculated) and the matched accepted count is calculated (recalculated) for 5000 resamples. Seed 20261113 fixes resampling. Exploratory 95% percentile intervals are reported and zero-acceptance resamples are revealed. These intervals are based on the fitted threshold, and not all the uncertainty due to choice of development set or model is accounted for.

Four prespecified thresholds are reported without choosing a replacement with evaluation labels. The number of images that contributed to the accepted and the false positives and negatives direct answer, rejected correct positives and rejected false positives are reported. The runtime is used to distinguish between generator seconds/question and detector seconds/ image. Model loading and downloading are excluded from runtime measurements. The evidence -sufficiency annotations, calibration guarantees and causal image-intervention tests are not part of the executed evaluation.

6. Results

6.1. Exploratory Pilot:

In its 120 questions for which a direct answer was provided, there was an agreement with the reference on 69.2%. The positive gate received 21 responses with a 17.5% coverage and 95.2% agreement. The 95.2% result for positive only confidence selection was also for the same count. The gate retained the only reference-disagreeing positive answer while rejecting four correct positive answers. This pilot study proved that the execution was possible, but this study did not show that the pilot provided additional predictive value. The provenance audit also limited its evidential value, which was the reason to motivate the new validation-split cohort.

6.2. New Validation-Split Replication:

The new evaluation consists of 240 questions based on 40 images, with 120 positive and 120 negative question-level reference labels. There were 49 positive and 191 negative answers from direct generation with zero unparseable outputs. The false-positive count was 2, while the false-negative count was 73, with an overall rate of reference agreement being 68.75%.

The detector threshold used was 0.3 and the development target was achieved. The gate accepted 39 answers from 31 evaluation images, resulting in coverage of 16.25% of evaluation questions and 100.00% reference agreement. The reference-disagreeing positive count in this subset was 0. The all-answer confidence control obtained a score of 79.49%, and the positive-only confidence control a score of 97.44% at the same count. The gate had a higher accepted-answer accuracy point estimate than positive-only confidence selection, but the exploratory difference interval included zero; superiority over that control was not established.

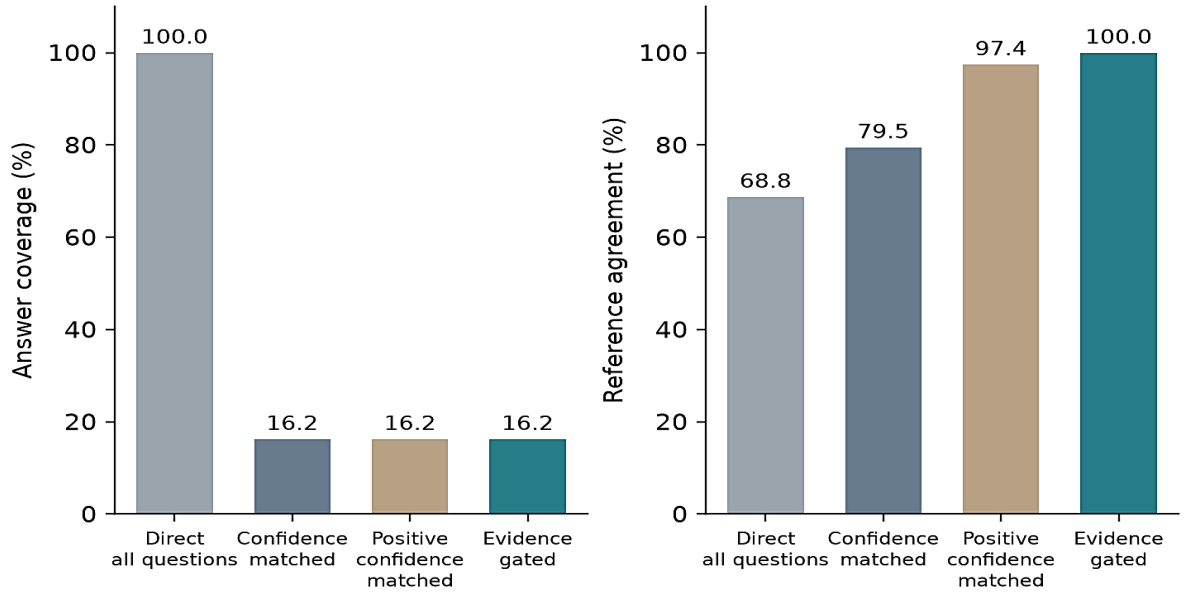


Figure (1): Measured coverage and reference agreement in the validation-split cohort. Selective conditions accept the same number of questions. Matching answer-class composition is necessary to interpret the gate's apparent accuracy relative to full direct answering.

The exploratory 95% paired bootstrap interval for the gate-minus-positive-control accuracy difference was [0.0, 8.6] percentage points. For the gate-minus-all-answer-control difference, it was [10.5, 30.8] percentage points. Out of 5000 resamples, there were 0 excluded resamples with zero accepted answers. The interval which is degenerate or narrow in a small sample should not be interpreted as a population guarantee, especially in the case of selection procedures which keep the same incorrect examples.

6.3. Positive-Answer Retention and Sensitivity:

In the full direct-positive pool 49 answers were obtained and 95.92% agreement was reached. The gate was able to return 39 correct positives, and reject 8 correct positives and 2 false positives. These counts separate the number of rejected presence claims which were wrong from the number of correct claims that were selected from a more limited set.

The gate was accepted 39 times at threshold 0.3, and at 100.00% agreement and 16.25% coverage. 38 answers were received at a 15.83% coverage level with 100.00% agreement at threshold 0.5. The 14.58% coverage and 100.00% agreement were obtained at threshold 0.7 with 35 answers accepted at the gate. 31 answers were accepted by this gate at threshold 0.9 with 12.92% coverage and 100.00% agreements.

These sensitivity measures refer to the set of thresholds pre-specified. They were not used to replace the development selected primary operating point. Interpretation of small percentage change is restricted by different subsets accepted and by the different images included in each subset.

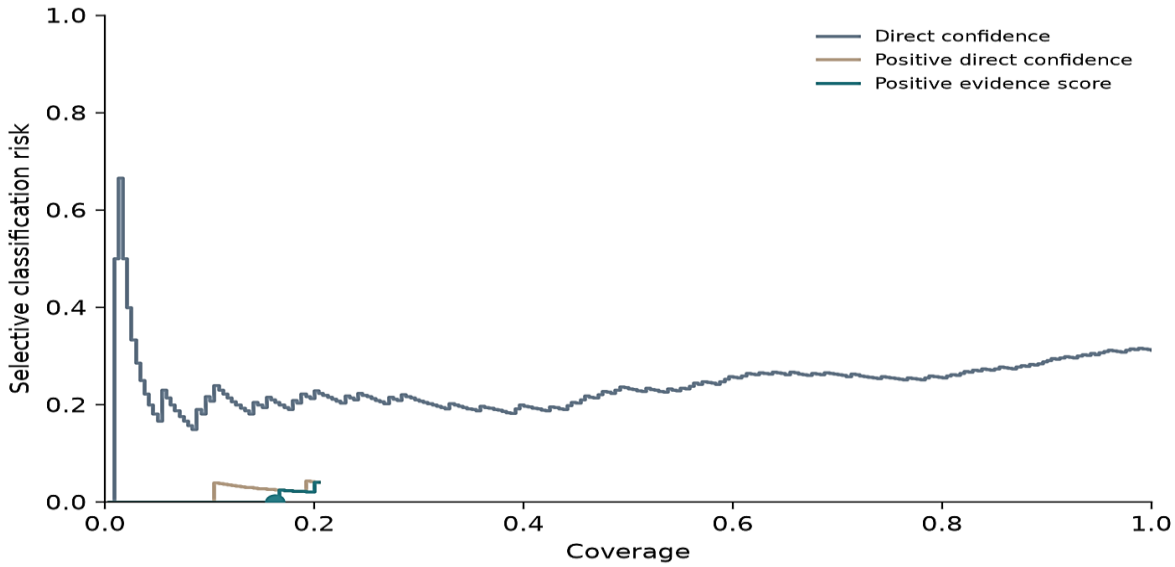


Figure (2): Validation-split cohort exploratory risk coverage rankings. The frozen development threshold is used in the marked gate point. The positive-only curves do not extend to full coverage because negative candidates always abstain.

6.4. Computational Cost:

The median processing time for the CPU configuration was 4.20 seconds per evaluation question with a 95th percentile of 5.10 seconds per evaluation question. The median computation time for the detector was found to be 0.21 s per evaluation image. The total evaluation generator processing time was 1011.5 seconds and the processing time spent on the detector was 9.1 seconds. The evidence from the detectors was broadcasted over six questions for each image. The additional cost of the gate

must be taken into account based on the accuracy of the matched control and not be compared with an unmatched control distribution.

7. Discussion

The main question is whether there is a difference between auditability and predictive benefit. Up to now, the acceptance decision is possible to inspect only with saved detector boxes, which do not automatically give semantic correctness for the answer. A box can be the wrong one, be mislabelled or insufficient to support the requested interpretation. This study is therefore based on reference agreement as the outcome and will not make it "verified grounding" or "causal visual faithfulness."

There are many comparison variables, but one of the most significant is answer class composition. A generator with different error rates for positive and negative answers can appear substantially more accurate after a policy accepts only positives. Full-coverage direct accuracy is therefore not the appropriate primary reference for such a gate. Positive-only matched-count control provides a more focused comparison and positive-retention counts indicate actual changes in the external detector. This does not mean that all evidence-based acceptance systems are without value; the conclusion is limited to the design being evaluated.

The same goes for the origin of the components. Separating development and evaluation images within a new experiment does not remove a pretrained detector's earlier training exposure. The pilot audit revealed that distinction and the replication was kept under control by means of official validation-image identifiers. But even with the use of the validation-split membership, independence from all model development or generator pretraining is not guaranteed. Stronger generalization claims could be made with the use of an external sequestered dataset with documented provenance.

7.1. Limitations:

The experiments employ a single small generator and a single detector, a single question category, and a single adversarial benchmark subset. The replication is a small sample of the full benchmark, not the actual benchmark. The first-token probability (confidence heuristic) can be less effective than learned selectors or sample based uncertainty estimators. Thus, evidence-gated answering is only compared to explicit lightweight controls and not all possible selective VQA methods.

The gate cannot accept negative answers, and does not perform general multimodal reasoning. Its coverage is structurally limited by the generator's positive predictions and the detector's vocabulary and recall. Not an independent human annotator judged the relevance of boxes, nor was there any counterfactual masking to determine whether cited pixels caused an answer. The bootstrap retains the dependence among questions within a given image, but does not account for any lack of image diversity or errors in the labels assigned to the datasets or uncertainty with the model exposed to the images. It is exploratory in nature and is not to be interpreted as confirmation of the equivalence or superiority of this work.

7.2. Recommendations and Future Work:

Evidence-constrained VQA studies should include accuracy as well as accepted coverage, use answer-class-matched controls, and include the disclosure of the provenance of the training data of all external components. Reference-label agreement and evidence sufficiency judgement from human must be evaluated separately. Future experiments will allow comparisons to be made of larger generators, calibrated selection scores, multilingual questions, and independently annotated regions. It is important to have a specific absence model and uncertainty model for reliable absence verification instead of assuming absence when the detection is not received.

Atomic claims, labels supported/contradicted/insufficient and selective acceptance are optional parts of the broader EC-VQA concept, which is future work here. This case study should not be used as an endorsement of that overall design. Extension should first be demonstrated to be incrementally better when compared to simpler matched controls, with a prospectively agreed protocol.

8. Conclusions

A reproducible object-presence evidence gate for visual question answering is evaluated. It maintains localised detector information, analyzes the outcomes of various queries by coverage and answer-class composition and performs audits of the provenance of component images. The independent-image replication evaluates 240 questions from 40 COCO val2017 images after development on a separate image subset.

The gate accepted 39 answers with 100.00% reference agreement at 16.25% coverage; positive-only confidence selection at the same accepted count achieved 97.44% reference agreement. The gate had a higher accepted-answer accuracy point estimate, but the exploratory difference interval included zero; superiority over positive-only confidence selection was not established. The results that are found here are to constrict the utility of the evaluated lightweight gate and to show that a bounding-box record combined with high selected-subset accuracy does not establish general hallucination mitigation.

The contribution consists of an open and honest diagnostic evaluation along with concrete outputs and clear limitations. For broader claims, there will be more than one model, independent evidence annotations, and more control of training exposure. Negative or null comparisons are included as it identifies which parts of the evidence-constrained answering claim are supported by measurement.

Declarations

Conflict of interest: The authors have no conflict of interest.

Availability of data and code: Data and code files are made available in the accompanying supplementary file. The compact archive does not contain model weights or original images, which may be downloaded with the preparation code provided, subject to their original licenses.

Ethics and writing help: No new study on human subjects or independent human annotation was performed. Using AI tools, the manuscript was drafted for writing, coding, and analysis operations. The authors are solely responsible for the technical content, the disclosure and final approval and for complying with the disclosure requirements of the Journal.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question Answering. Proceedings of ICCV. Primary paper.
2. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. Proceedings of CVPR, 6904–6913. Primary paper.
3. Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018). VizWiz Grand Challenge: Answering Visual Questions from Blind People. Proceedings of CVPR, 3608–3617. Primary paper.
4. Hudson, D. A., & Manning, C. D. (2019). GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. Proceedings of CVPR. Primary paper.
5. Radford, A., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. Proceedings of ICML, PMLR 139, 8748–8763. Primary paper.
6. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating Object Hallucination in Large Vision-Language Models. Proceedings of EMNLP, 292–305. Primary paper and DOI.
7. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. Advances in Neural Information Processing Systems, 36. Primary paper.
8. Guan, T., et al. (2024). Hallusion Bench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. Proceedings of CVPR, 14375–14385. Primary paper.
9. Yin, H., Si, G., & Wang, Z. (2025). Clear Sight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models. Proceedings of CVPR, 14625–14634. Primary paper.
10. Jo, Y., Bae, S., & Kim, T. (2026). Attention-Space Contrastive Guidance for Efficient Hallucination Mitigation in LVLMs. Proceedings of CVPR Findings, 9706–9715. Primary paper.
11. Marafioti, A., et al. (2025). SmolVLM: Redefining small and efficient multimodal models. arXiv:2504.05299. Technical report. Model card.

12. PyTorch Contributors. Faster R-CNN MobileNet V3 Large 320 FPN. Torchvision documentation. Model and weight specification.
13. LMMs-Lab-Encoder. POPE dataset mirror. Hugging Face Datasets. Original benchmark: Li et al. [6]. Dataset source.
14. COCO Organizers. (2017). COCO 2017 Detection Overview. Official workshop presentation. Train and validation split reorganization.
15. Whitehead, S., Petryk, S., Shakib, V., Gonzalez, J., Darrell, T., Rohrbach, A., & Rohrbach, M. (2022). Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly. Proceedings of ECCV. Primary paper.
16. Eisenschlos, J., Maina, H., Ivetta, G., & Benotti, L. (2024). Selectively Answering Visual Questions. Findings of ACL, 4219–4229. Primary paper and DOI.
17. Srinivasan, T., Hessel, J., Gupta, T., Lin, B. Y., Choi, Y., Thomason, J., & Chandu, K. (2024). Selective "Selective Prediction": Reducing Unnecessary Abstention in Vision-Language Reasoning. Findings of ACL, 12935–12948. Primary paper and DOI.
18. Kotte, V. (2026). EVICT: Evidence-Sufficiency Verification via Counterfactual Dropout for Visually-Grounded Selective Question Answering. Proceedings of CVPR Workshops, 11251–11259. Primary paper.