UnbiasedVQA: Toward Unbiased Benchmark for Assessing General Scene Understanding of Multimodal Large Language Models

Anonymous ACL submission

Abstract

Recent multimodal large language models (MLLMs) demonstrate strong scene understanding, yet often produce confident hallucinations, making reliable evaluation essential. Existing general scene understanding visual question answering (VQA) benchmarks are often COCO-based, yet COCO is widely exposed in MLLM training pipelines, directly or via derivative datasets. Through statistical analysis and controlled experiments, we observe that such non-negligible exposure biases model's generalization evaluation, distorting comparisons of both individual MLLMs and MLLM-fusions. We introduce UnbiasedVQA [1], COCO-free benchmark for unbiased assessment of MLLM's general scene understanding. UnbiasedVQA is a multi-choice VQA, built from images statistically verified to be unseen during MLLM training. We carefully design the benchmark to improve evaluation quality through broader question diversity, more challenging visual understanding tasks, uncertainty-aware metric and carefully verified human annotations. Our extensive experiments show that UnbiasedVQA supports rigorous evaluation of both standalone MLLMs and supervised or unsupervised MLLM fusion systems, while COCO-based benchmarks can yield misleading conclusions.

1 Introduction

Recent multimodal large language models (MLLMs) have demonstrated impressive capabilities in visual question answering (VQA) in general scene understanding. They, however, still produce incorrect answers with high confidence, particularly when visual evidence is ambiguous: for example, when answering requires attention to non-salient regions or to small or occluded objects. Current training and evaluation practices often reward confident guessing rather than calibrated uncertainty (Kalai et al., 2025). As a result, when faced with difficult or ambiguous questions, models may generate plausible but wrong answers instead of abstention, even in state-of-the-art systems. To mitigate this issue, prior efforts have focused on improving individual MLLMs or combining multiple MLLMs through model fusion (Li et al., 2023b).

Reliable evaluation is essential for such progress. A benchmark should not only measure whether an MLLM answers correctly with right confidence, but also whether the evaluation itself reliably reflects the model's true generalization ability rather than familiarity with images, annotations, or biases inherited from benchmark-related training data. Numerous VQA benchmarks have been developed to assess different capabilities of MLLMs. Zhang et al. (2025a) briefly classifies VQA benchmarks into four categories: general scene understanding, reasoning, document understanding and OCR.

Despite their importance, general scene understanding VQA benchmarks are often derived from COCO (Lin et al., 2014). This raises a critical bias issue. COCO and COCO-derived datasets are non-negligibly exposed in the training pipelines of many MLLMs, either directly or indirectly. As a result, evaluation on COCO-based benchmarks may reflect MLLM's training-data familiarity rather than genuine generalization. In this paper, we substantiate this concern through both statistical analysis and controlled experiments. In particular, we find that several MLLMs exhibit a statistically significant performance difference between train and test splits of prior COCO-based benchmarks such as Visual7W (Zhu et al., 2016) and A-OKVQA (Schwenk et al., 2022). These results indicate that certain models may be biased toward COCO-based data, which can skew comparisons of model's generalization capability. This issue becomes even more consequential when evaluat-

[1] https://huggingface.co/datasets/anonymous-emnlp26/UnbiasedVQA

ing systems that combine multiple MLLMs. If a benchmark favors certain constituent models due to data familiarity, fusion results can be misleading. A fused system may over-weight models with benchmark-specific advantages rather than those with stronger visual scene understanding for a given sample, potentially degrading overall performance. We also demonstrate this problem on COCO-based benchmarks through controlled fusion experiments (See more details in Section 3.1).

To address these limitations, we propose UnbiasedVQA, a COCO-free benchmark for an unbiased assessment of general scene understanding in MLLMs. UnbiasedVQA is multi-choice VQA, constructed from images that are statistically verified to be unseen in MLLM training, enabling a more reliable model generalization assess. UnbiasedVQA is also designed to provide broader question diversity and more challenging visual understanding problems. In particular, questions are drafted using *text-only inputs* rather than, generated through image-conditioned process, encouraging questions that involve diverse concepts and non-salient objects, instead of the most obvious regions in the image. 10-option choice VQA lowers the chance of accidental correct answers compared with common multiple-choice formats that use five or fewer options. Beyond standard accuracy, we propose uncertainty-aware metrics for multi-choice VQA, Expected Calibration Error (ECE), to evaluate both correctness and confidence. This offers more informative characters of model behavior under hallucination risk, distinguishing models that are not only accurate, but also reliably calibrated. For rigorous ground truth answer, each example is further manually verified at the end of the data generation pipeline. UnbiasedVQA includes an optional training set, supporting evaluation of standalone MLLMs as well as supervised and unsupervised MLLM fusion systems. Through extensive experiments, we show that COCO-based benchmarks can lead to biased conclusions, while UnbiasedVQA provides more reliable assess of general scene understanding in MLLMs.

Our contributions are summarized below:

- We statistically reveal non-negligible MLLM bias on COCO-based VQA benchmarks and show that this bias propagates to MLLM fusion experiments, hindering trustworthy evaluation. To address this issue, we introduce UnbiasedVQA, a COCO-free benchmark that is statistically unbiased to models.
- UnbiasedVQA improves evaluation quality through more challenging questions involving non-salient objects that more likely trigger MLLM hallucinations, a 10-choice VQA format, uncertainty-aware metrics, and accurate ground truth ensured by final-stage manual verification.
- Our extensive experiments show that UnbiasedVQA enables a rigorous evaluation of both standalone MLLM and supervised or unsupervised MLLM fusion systems, while COCO-based benchmarks can lead to misleading conclusions.

2 Related Work

Recent multimodal large language models (MLLMs) have extended large language models with visual perception capabilities, enabling visual question answering, image captioning, grounding, and general scene understanding. Existing MLLMs span both open-source or open-weight models(Liu et al., 2023; Bai et al., 2023; Zhu et al., 2025; Lu et al., 2024; Li et al., 2023a; Dai et al., 2023; Chen et al., 2024a; Awadalla et al., 2023; Ye et al., 2024; Young et al., 2024; Bavishi et al., 2023; Ge et al., 2024) and commercial proprietary models(Hurst et al., 2024; Anthropic, 2024; Team et al., 2023). Despite these advances, MLLMs still suffer from reliability issues, including hallucinated responses, overconfidence on incorrect predictions, and instability to ambiguous or fine-grained visual evidence (Li et al., 2023c; Guan et al., 2024).

Since individual MLLMs have different strengths and weaknesses, model fusion has been explored as a system-level strategy to improve robustness and generalization. Classical ensemble learning aggregates multiple predictors (Hansen and Salamon, 2002; Breiman, 1996), while recent deep model fusion aggregates or routes across neural models to exploit their complementary strengths (Li et al., 2023b). In vision-language systems, MoE-LLaVA activates top-k experts through internal routers for sparse LVLM scaling (Lin et al., 2026), and MoME combines internal vision and language experts to mitigate task interference (Shen et al., 2024). Departing from such internal sub-network routing, recent advances have introduced model-level Mixture-of-Experts (MoE) frameworks that dynamically route inputs

across multiple independent, heterogeneous VLMs to mitigate hallucinations and capture visual uncertainty (Tekin et al., 2026; Zhang et al., 2026, 2025b). Although these studies suggest that model-level fusion is a promising direction for building more reliable MLLM systems, a router trained on data biased toward particular models may learn benchmark-induced expert preferences rather than generalizable model selection, thereby weakening the reliability of model fusion.

A wide spectrum of Visual Question Answering (VQA) benchmarks is developed to measure visual understanding and reasoning ability of MLLMs. Existing benchmarks are categorized into four groups based on their primary objectives: general scene understanding, reasoning, OCR, and document understanding. General scene understanding benchmarks (Goyal et al., 2017; Marino et al., 2019; Schwenk et al., 2022; Yu et al., 2023; Gurari et al., 2018; Chen et al., 2024b; Li et al., 2024) mainly evaluate visual perception, object recognition, and common sense reasoning in natural images. Reasoning benchmarks (Hudson and Manning, 2019; Saikh et al., 2022; Lu et al., 2023; Wang et al., 2024; Yue et al., 2025; xAI, 2024) emphasize multi-step inference, logical deduction, and domain-specific problem solving. These datasets require compositional reasoning over both visual and textual cues, often involving knowledge that extends beyond image content. OCR and text understanding benchmarks (Mishra et al., 2019; Singh et al., 2019) evaluate the model's ability to recognize embedded text within images. Document understanding benchmarks (Mathew et al., 2021, 2022; Masry et al., 2022; Kim et al., 2024) focus on the model's understanding of structured data, where spatial layouts, textual content, and visual reasoning are required. As argued by Tu et al. (2025), some benchmarks may have been exposed during MLLM training which degenerates trustworthy model assessment. We statistically verify the extent of this data-familiarity issue and introduce a new benchmark designed to mitigate it, enabling more reliable comparisons of genuine model generalization.

3 Limitations and proposals

We statistically show that COCO-based VQA benchmarks exhibit non-negligible bias toward certain models, leading to unreliable generalization evaluation of individual MLLM and fused models. To address these limitations, we propose Unbiased-VQA, a COCO-free benchmark designed to reduce dataset-specific bias and enable trustworthy assessment of general scene understanding. Unbiased-VQA also provides improved benchmark properties, including more challenging visual understanding tasks and Uncertainty-aware metric.

3.1 Removal of non-negligible model bias on COCO-based benchmarks

VQA benchmarks for general scene understanding are often built on COCO images (Lin et al., 2014) which are also commonly included in the pretraining corpora of modern MLLMs. For instance A-OKVQA (Schwenk et al., 2022), Visual7W (Zhu et al., 2016), and VQAv2 (Marino et al., 2019) VQA benchmarks are originated from COCO. And, Qwen-VL (Bai et al., 2023), InternVL3 (Zhu et al., 2025), Phi-4 (Abdin et al., 2024), and PLM (Cho et al., 2025) are reportedly pretrained on large-scale heterogeneous data, with some sources overlapping with (or derived from) COCO. This overlap raises concerns about benchmark-specific bias and the reliability of model comparisons: some models may obtain inflated benchmark accuracy due to prior exposure to similar image-text pairs during pretraining, rather than demonstrating genuine visual understanding and robust generalization. Whereas, UnbiasedVQA is constructed with COCO-free data and statistically ensure that the training data is less likely observed in training MLLMs.

To quantify this assertion, we test whether the train-test accuracy gaps of MLLMs on existing benchmarks are statistically significant using a two-proportion z-test. Under the null hypothesis $H_0 : \hat{p}_{\text{train}} = \hat{p}_{\text{test}}$, where $\hat{p}_{\text{train}}$ and $\hat{p}_{\text{test}}$ denote accuracies of a MLLM on the training and test splits, we compute the corresponding p-value (Fisher, 1970). A smaller p-value indicates that the model accuracy discrepancy between the splits is statistically significant which means that model is highly likely exposed to train dataset.

Figure 1 show that multiple MLLMs achieve statistically significant higher accuracy on the train split than on the test split on A-OKVQA and Visual7W [2]. On A-OKVQA, Phi-4 (Abdin et al., 2024) exhibits the largest train–test gap (about 13%), and other models such as LLaVA1.6 and InternVL3 show similarly pronounced discrepancies with low p-values. Whereas, GPT-4o presents

[2]The Telling subset is used for all the experiments unless otherwise specified.

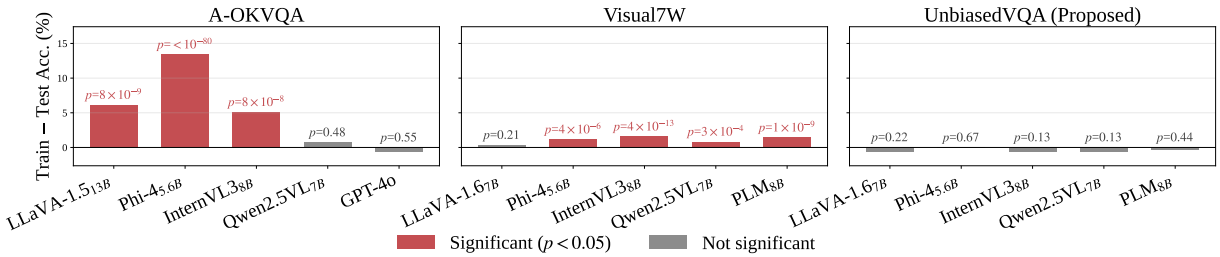


Figure 1: **Comparison of Train–Test acc. and p-value in z test across COCO-based and UnbiasedVQA benchmarks:** A smaller p-value indicates that the accuracy discrepancy between the two splits is more statistically significant. A-OKVQA and Visual7W exhibit statistically significant performance gaps for several MLLMs, suggesting the presence of dataset bias in COCO-based benchmarks. In contrast, UnbiasedVQA shows no statistically significant train–test differences across the evaluated models, indicating that it provides a more reliable benchmark for assessing general scene understanding.

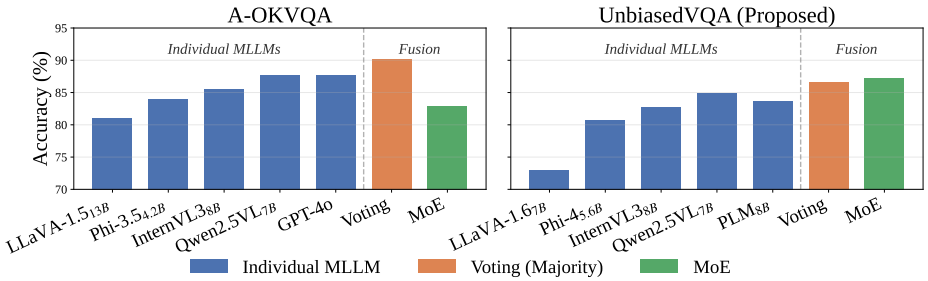


Figure 2: **Individual and fusion MLLM accuracy comparison on A-OKVQA and UnbiasedVQA:** When the MoE (Shazeer et al., 2017) is evaluated on COCO-based benchmark, A-OKVQA, its accuracy drops substantially. MLLMs that have been directly or indirectly exposed to COCO during training, such as Phi, tends to perform relatively better than other models on A-OKVQA. As a result, the MoE assigns higher weights to these models not because they demonstrate stronger general VQA capability, but because they are more familiar with the benchmark distribution. Such biased model selection degrades the overall fusion performance. Whereas, the same MoE approach outperforms the other methods on UnbiasedVQA.

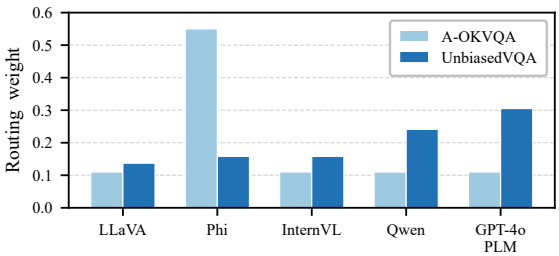


Figure 3: **Distribution of MoE (Shazeer et al., 2017) routing weights when trained on A-OKVQA versus UnbiasedVQA train set:** When trained on A-OKVQA, the router assigns disproportionately high weights to Phi likely due to its familiarity with the COCO-based benchmark distribution, not genuine image-understanding capability which degrades MoE performance.

higher p-value, indicating no statistically meaningful accuracy discrepancies between train and test sets. On Visual7W, Phi-4, Qwen2.5VL, and InternVL3 also yield low p-values, indicating that this pattern persists even in a larger COCO-based benchmark. The consistent and statistically significant advantage on the training split across both datasets suggests that these benchmarks weaken evaluation reliability, and are insufficient to measure model's generalization capability. In contrast, Figure 1 reports substantially higher p-values across all the models on UnbiasedVQA, implying smaller and less systematic train-test accuracy differences. This suggests that proposed UnbiasedVQA is less revealed in MLLM's pretraining and thus provides more reliable assessment of MLLMs' generalization capability than COCO-based benchmarks such as A-OKVQA and Visual7W.

Controlled MLLM fusion experiments further demonstrate that biases in COCO-based benchmarks can lead to distorted evaluations of model fusion. MLLM fusion provides a natural strategy to mitigate individual model failures by leveraging complementary strengths across multiple models. Our controlled experiments, however, show that bias in COCO-based benchmark can also propagate to, and even degrade, fused systems. To examine this, we train a Mixture-of-Experts (MoE) framework (Shazeer et al., 2017) with A-OKVQA and

Table 1: **UnbiasedVQA noun diversity and bbox size of queried objects:** When questions are drafted based on text, the noun diversity is increased and bbox size of queried objects is smaller, which makes VQA more diverse and challenging.

Input to GPT	Total Nouns	Unique Noun	Median of max bbox
Only Text	46,047	2,276	0.2526
Image + Text	42,369	2,163	0.2697
Only Image	39,606	887	0.3062

Table 2: **5-choice VS 10-choice VQA accuracy:** Accuracies on 10-choice VQA is consistently lower than typical 5-choice VQA.

Model Size	LLaVA1.6 7B	Phi-4 5.6B	InternVL3 8B	Qwen2.5VL 7B	PLM-8B 8B
5-choice	80.65	86.22	88.62	87.22	87.85
10-choice	71.60	78.91	79.51	78.71	80.18

UnbiasedVQA and learn MoE router on the training data. As shown in Figure 2, When trained on A-OKVQA, its performance drops by nearly -10% compared with the simple voting baseline. MLLMs that have been directly or indirectly exposed to COCO during training, such as Phi, tend to perform relatively better on A-OKVQA than other models. Consequently, the MoE assigns larger routing weights to these models not because they possess stronger general VQA capability, but because they are more familiar with the benchmark distribution. This biased expert selection ultimately harms the overall fusion performance. Figure 3 visualizes this effect: the MoE trained on A-OKVQA produces highly skewed routing weights. This indicates that COCO-based benchmarks can cause the fusion model to overfit to dataset-specific bias rather than learn generalizable expert selection. In contrast, the MoE trained on UnbiasedVQA achieves a significant performance gain of $+3.67\%$, outperforming both individual models and unsupervised voting with more balanced expert-selection distribution as shown in Figure 3.

3.2 More diverse and challenging VQA

UnbiasedVQA provides a more diverse and challenging VQA benchmark. Our VQA generation pipeline increases noun diversity in questions, raises difficulty by focusing more on non-salient objects, and adopts a 10-choice VQA rather than the typical (5 or fewer) options.

UnbiasedVQA initializes questions and candidate answers with GPT using only text descriptions as input; i.e., images are not used when drafting questions. When images are included when drafting question with models, the question can be implicitly shaped by the model's pretrained image–text alignment priors, which may constrain word choice and reduce linguistic and semantic diversity. To mitigate this distribution-induced bias, UnbiasedVQA employs a text-only generation strategy. Conditioning exclusively on texts decouples question formulation from the generator's internal visual priors, enabling more flexible word selection and yielding substantially richer question phrasing, reasoning patterns, and visual concepts. Table 1 supports this, showing that text-only input achieves higher noun diversity of questions than image-only and image+text input.

Using only text descriptions as input also encourages questions related to non-salient objects because, if images are given, GPT likely attends to the most visually prominent regions while drafting questions. Table 1 shows that the object bounding boxes associated to the questions are smaller when only using text as an input to GPT, indirectly indicating a stronger focus on non-major objects. This increases difficulty for MLLMs because uncertainty often rises when the queried object occupies a small image region (Luo et al., 2024; Chen et al., 2025). See more examples in Figure 5.

For answer construction, UnbiasedVQA uses a 10-choice format for each question. Increasing the number of candidate answers enables more diverse and harder distraction while reducing the chance of correct guessing, leading to a more reliable evaluation. As shown in Table 2, the consistent accuracy drop across MLLMs validates that a larger answer space yields a more rigorous assessment. This design improves answer diversity and supports a more robust evaluation of visual grounding and uncertainty handling.

3.3 Uncertainty-aware metrics on Multi-choice VQA

UnbiasedVQA further incorporates uncertainty-aware evaluation metrics operating on the multiple-choice VQA format (Pandey et al., 2026). Compared with open-ended VQA, multiple-choice VQA restricts responses to a predefined set of candidate answers, which enables more objective, reproducible, and scalable evaluation by simplifying answer validation and reducing reliance on subjective semantic matching (Zhang et al., 2025a). Although accuracy provides a direct measure of task performance, it does not capture whether a

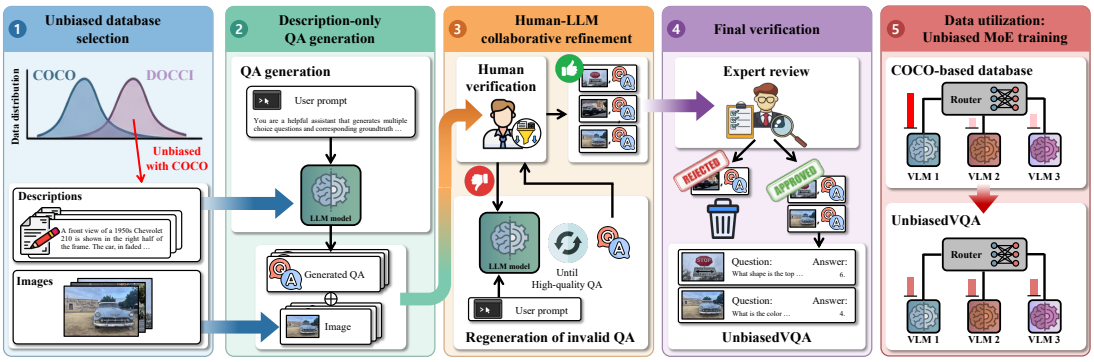


Figure 4: **Overall pipeline for constructing UnbiasedVQA and its downstream utilization:** The dataset is constructed through COCO-free dataset, description-only QA generation, and human-LLM collaborative refinement. The pipeline ensures the groundtruth reliability with human expert verification at the final stage.

Q) What text is visible on the clock in the image?

Q) What is the color pattern of the wine glass on the far left?

Q) What is visible on the pavement around the car?

Q) How many tree vines are hanging from the roof of the building?

Figure 5: **Examples of UnbiasedVQA:** Red boxes indicate objects associated to the questions in UnbiasedVQA. Questions are more related to non-salient objects where MLLMs likely present hallucination.

model is adequately confident in its predictions, which is critical for reliable assessment. Since UnbiasedVQA is formulated as a multiple-choice benchmark, confidence-aware metrics such as Expected Calibration Error (ECE) can be naturally incorporated to evaluate not only correctness, but also calibration quality. More details are provided in Appendix C.

4 UnbiasedVQA Benchmark

4.1 Data generation pipeline

Figure 4 illustrates the overall pipeline for constructing UnbiasedVQA. It is built on the COCO-free Docci dataset (Onoe et al., 2024), which serves as an unbiased database source in our construction pipeline and provides images paired with dense, object-centric, and relation-aware textual descriptions. We draft multiple-choice questions and answers via GPT-4o (Hurst et al., 2024) where the input is only the high-fidelity text descriptions, instead of typical image or image + text. This setting increases relibility, diversity and difficulty of questions. Refer to Section 3 for more details.

To draft multiple-choice questions, we adopt a prompting strategy based on (Zhang et al., 2024) to steer GPT-4o toward generating diverse question types. For each image, we generate three QA pairs, with prompts encouraging broad coverage of visual semantics, including global scene content, object attributes, human appearance, spatial layout, contextual cues, and observable interactions. The prompt also permits questions involving counting, spatial relations, and camera perspective when such information is explicitly supported by the text description. Throughout generation, each question is constrained to remain grounded in the provided description and to have a single, unambiguous correct answer among the options.

To ensure data ground-truth reliability, all generated QA pairs undergo manual verification by ten expert-human annotators. Each annotator evaluates the data according to strict criteria, discarding samples with incorrect or ambiguous labels, reliance on external knowledge, or poor visual grounding. When human verification indicates low generation quality in the QA dataset, we revise the prompts based on the human feedback and regenerate the dataset using GPT-4o. This process is repeated until a high-quality QA set is obtained. After this human–LLM collaborative refinement, expert reviewers conduct a final verification step and discard any remaining unreliable samples. The annotation process involves no personal data and poses no known risks to participants. This human-in-the-loop process effectively refines the dataset by

Table 3: **Statistics of general scene understanding multiple-choice VQA benchmarks**

	QA	Images	Questions/Image	Choices/Question	Avg.QLen	Avg.CLen	COCO
BEAF	26, 064	2, 223	11.72	2	7.17	1	O
A-OKVQA	24, 903	23, 692	1.05	4	8.82	1.24	O
Visual7w	139, 868	28, 653	4.88	4	5.64	1.90	O
UnbiasedVQA(ours)	36, 241	14, 460	2.51	10	9.84	3.32	X

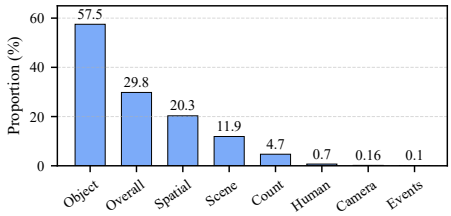


Figure 6: **Distribution of QA types in UnbiasedVQA**

eliminating generation-induced errors, establishing UnbiasedVQA as a highly reliable benchmark.

4.2 Data statistics

The UnbiasedVQA consists of 36,241 multiple-choice question–answer pairs constructed from 14,460 images. They are divided into 23,905 training samples and 12,336 test samples, enabling evaluation under both standalone MLLM settings and supervised or unsupervised fusion frameworks. Unlike several existing COCO-based benchmarks that exhibit significant train–test discrepancies across MLLMs, UnbiasedVQA maintains relatively consistent train–test performance across evaluated models (Figure 1). This stable benchmark distribution enables not only more reliable generalization evaluation, but also meaningful comparison of supervised fusion approaches such as Mixture-of-Experts (MoE). Each question is associated with 10 answer choices, making UnbiasedVQA a large-answer-space VQA benchmark. As summarized in Table 3, UnbiasedVQA features marginally longer questions and answer choices on average. These statistics directly reflect the prompt used during data generation encourages questions to describe scene-level visual content rather than ask isolated attributes. By covering diverse aspects such as overall image content, objects and their attributes, spatial layout, and scene context, the prompt produces questions that require comprehensive understanding of how visual elements are organized within a scene. Figure 6 shows the distribution of QA types in UnbiasedVQA. Each question may be associated with multiple types, reflecting that UnbiasedVQA questions often capture several aspects of a scene simultaneously. While object-related content appears frequently, many questions also involve overall image understanding, spatial relationships, and scene context, indicating that UnbiasedVQA emphasizes holistic scene understanding rather than isolated object recognition.

Figure 5 presents examples of objects associated with the question in UnbiasedVQA. Rather than relying only on salient objects, the questions require attention to peripheral and scene-level elements such as background text, subtle object attributes, pavement details, and small structures within the environment. These examples highlight the dataset's emphasis on fine-grained and holistic visual understanding.

5 Experiments

Experiment Setting: We compared UnbiasedVQA with COCO-based benchmarks using a diverse set of pretrained MLLMs, including LLaVA (Liu et al., 2023), Phi-4 (Abdin et al., 2024), InternVL3 (Zhu et al., 2025), Qwen2.5VL (Bai et al., 2025), PLM (Cho et al., 2025), and GPT-4o (Hurst et al., 2024). In addition to evaluating individual models, we further investigate ensemble-based approaches composed of multiple MLLMs. Specifically, we consider two ensemble baselines: (1) an unsupervised majority voting strategy that aggregates predictions from multiple models, and (2) a supervised Mixture-of-Experts (MoE) framework (Shazeer et al., 2017), where each pretrained MLLM is treated as a frozen expert and a lightweight router learns to combine expert predictions conditioned on the image-question pair. The availability of a dedicated training split in UnbiasedVQA enables supervised training of such routing-based ensemble methods in addition to standard evaluation of standalone models.

Statistical Bias Test: To investigate whether existing VQA benchmarks exhibit dataset-specific bias or train–test inconsistency, we first compare train–test accuracy gaps across multiple benchmarks in Figure 1, with detailed numerical results provided in Appendix D. Several MLLMs show statistically significant discrepancies on COCO-based datasets such as A-OKVQA and Visual7W, reflected by low p-values and noticeable train–test

Table 4: **Individual and fusion MLLM accuracy on BEAF, A-OKVQA, and UnbiasedVQA:** MLLMs present lower accuracy on UnbiasedVQA due to its challenging questions. Note that BEAF does not provide training dataset, unable to run MoE experiments.

Dataset		Individual MLLM Acc.					Fusion Acc.	
		1	2	3	4	5	Voting	MoE
BEAF	Model	LLaVA-1.5$_{13B}$	GPT-4o	Phi-3.5$_{4.2B}$	Qwen2.5VL$_{7B}$	InternVL3$_{8B}$	Majority	MoE
	Acc.	83.18	86.82	86.93	87.88	89.28	**89.47**	N/A
A-OKVQA	Model	LLaVA-1.5$_{13B}$	GPT-4o	Phi-3.5$_{4.2B}$	Qwen2.5VL$_{7B}$	InternVL3$_{8B}$	Majority	MoE
	Acc.	81.00	87.69	83.94	87.69	85.46	**90.19**	82.88
UnbiasedVQA (Proposed)	Model	LLaVA-1.6$_{7B}$	Phi-4$_{5.6B}$	InternVL3$_{8B}$	Qwen2.5VL$_{7B}$	PLM$_{8B}$	Majority	MoE
	Acc.	72.92	80.61	82.76	84.86	83.56	86.54	**87.18**

Table 5: **Performance on UnbiasedVQA**: Accuracy and ECE of each MLLM, a unsupervised majority voting and a supervised Mixture-of-Experts (MoE) (Shazeer et al., 2017)

Model size	LLaVA-1.6 7B	Phi-4 5.6B	InternVL3 8B	Qwen2.5VL 7B	PLM 8B	Voting	MoE
Acc.↑	72.92	80.61	82.76	84.86	83.56	86.54	87.18
ECE ↓	0.1003	0.0917	0.0479	0.2320	0.1216	0.0558	0.0695

accuracy gaps. In particular, models such as Phi-4 achieve substantially higher training accuracy than test accuracy, suggesting possible benchmark bias or distribution overlap. In contrast, UnbiasedVQA maintains stable p-values and highly consistent train/test accuracy across all evaluated MLLMs. These results suggest that UnbiasedVQA provides a more balanced evaluation setting that better reflects model generalization assessment rather than benchmark-specific memorization effects.

Controlled MLLM Fusion Experiment: We further evaluate how such benchmark characteristics affect ensemble-based learning. As shown in Figure 2 and Table 4, ensemble methods generally improve accuracy across datasets, but the effectiveness of supervised MoE training differs substantially depending on the benchmark distribution. On COCO-based datasets such as A-OKVQA, models with large train accuracy can dominate router optimization, making stable MoE training difficult. For example, experts with excessively high training accuracy may receive disproportionately large routing weights despite weaker generalization performance on the test set. In contrast, because UnbiasedVQA exhibits relatively consistent train/test behavior across all MLLMs, the MoE router can learn more balanced expert utilization, resulting in the best performance on UnbiasedVQA. These findings indicate that reliable benchmark distributions are important not only for unbiased evaluation, but also for stable supervised ensemble learning.

UnbiasedVQA Benchmark Result: We note that Table 4 shows that individual MLLM accuracies on UnbiasedVQA are generally lower than those on COCO-based benchmarks due to the design of our challenging VQA.

Table 5 presents the performance of individual MLLMs as well as two ensemble baselines, including unsupervised majority voting and a supervised Mixture-of-Experts (MoE) framework. In addition to accuracy, we report Expected Calibration Error (ECE) to evaluate whether model confidence reliably reflects prediction correctness. Among single models, Qwen2.5VL achieves the highest accuracy but with substantially higher ECE compared to other models, indicating that its confidence estimates are less reliable despite strong prediction performance. In contrast, the ensemble approaches achieve low ECE, while further improving overall accuracy. These results suggest that ensemble-based methods can enhance not only prediction correctness but also confidence score, highlighting the importance of confidence-aware evaluation beyond accuracy alone.

6 Conclusion

We introduce UnbiasedVQA, multi-choice VQA benchmark for trustworthy assessment of MLLMs' general scene understanding. Built from images statistically verified as unseen during training, it offers diverse questions, challenging visual tasks, uncertainty-aware evaluation, and carefully validated annotations. Through extensive experiments, we show that UnbiasedVQA enables more rigorous evaluation of both individual MLLMs and MLLM fusion systems, while COCO-based benchmarks can lead to misleading conclusions.

Limitations

While UnbiasedVQA provides a reliable benchmark for evaluating MLLMs and MLLM fusion methods, it has following limitations that point to directions for future work.

First, UnbiasedVQA is currently built on the Docci dataset, which limits the overall scale of the benchmark despite its high-quality descriptions. As a result, UnbiasedVQA is currently smaller in scale than large web-crawled datasets. Future work will expand the benchmark by incorporating web-crawled image–text pairs while preserving its COCO-free construction and safety-oriented evaluation design.

Second, UnbiasedVQA focuses on image VQA task. Although image VQA is a fundamental testbed for scene understanding and uncertainty-aware reasoning, many real-world safety-critical applications involve temporal and dynamic visual inputs. Extending UnbiasedVQA to video would allow evaluation of uncertainty, hallucination, and calibration under temporal ambiguity, motion blur, and long-range dependencies, which are not captured in static images.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sagnak Tasırlar. 2023. Fuyu-8b: A multimodal architecture for ai agents. *URL: https://www. adept. ai/blog/fuyu-8b*, 3.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2024a. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning.

Liangyu Chen, James Burgess, Jeffrey J Nirschl, Orr Zohar, and Serena Yeung-Levy. 2025. The impact of image resolution on biomedical multimodal large language models. *Preprint*, arXiv:2510.18304.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, and 1 others. 2025. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024. Making llama see and draw with seed tokenizer. In *International Conference on Learning Representations*, volume 2024, pages 35206–35231.

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

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Lars Kai Hansen and Peter Salamon. 2002. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. 2023b. Deep model fusion: A survey. *arXiv preprint arXiv:2309.15698*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. 2026. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, pages 1–14.

Tsung-Yi Lin, Michael Maire, Serge 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*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. 2024. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, and 1 others. 2024. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pages 291–309. Springer.

Ayush Pandey, Jai Bardhan, Ishita Jain, Ramya S Hebbalaguppe, Rohan Raju Dhanakshirur, and Lovekesh Vig. 2026. Refine and align: Confidence calibration

through multi-agent interaction in vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37810–37819.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. 2024. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Yahan Tu, Rui Hu, and Jitao Sang. 2025. Ode: Open-set evaluation of hallucinations in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19836–19845.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

xAI. 2024. Realworldqa dataset. https://x.ai/. Dataset release.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 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.

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

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, and 1 others. 2025a. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29580–29590.

Yulong Zhang, Tianyi Liang, Xinyue Huang, Erfei Cui, Xu Guo, Pei Chu, Chenhui Li, Ru Zhang, Wenhai Wang, and Gongshen Liu. 2025b. Consensus entropy: Harnessing multi-vlm agreement for self-verifying and self-improving ocr. *arXiv preprint arXiv:2504.11101*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 4995–5004.

A UnbiasedVQA example

After QA generation, ten annotators verified the samples following predefined guidelines. A QA pair was marked as invalid if it contained (1) an incorrect ground-truth answer, (2) multiple plausible correct answers, (3) a mismatch between the question and the answer, (4) reliance on specialized or external domain knowledge beyond the description, or (5) abnormal or ill-formed questions. For verification, the generated QA pairs were randomly and evenly split across annotators to ensure balanced workloads. Figures 7 and 8 present representative examples of retained samples and cases filtered out during this process.

B QA generation Detail

To ensure reproducibility, we provide the prompt used to generate multiple-choice QA pairs from Docci descriptions. The prompt is designed to encourage diverse coverage of scene-level visual aspects while enforcing a single correct answer among multiple choices.

C Metrics

Most existing evaluations of MLLMs rely on accuracy, which measures correctness but does not capture prediction confidence or uncertainty. This limitation becomes more pronounced in challenging settings such as large answer spaces (e.g., 10-choice QA), where reliable uncertainty estimation and calibration are essential. To address this gap, we complement accuracy with Expected Calibration Error (ECE), enabling a more fine-grained assessment of both discriminative ability and confidence reliability.

Expected Calibration Error measures the discrepancy between a model's predicted confidence and its empirical accuracy. Predictions are partitioned into M confidence bins $\{B_m\}_{m=1}^{M}$, and ECE is computed as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \quad (1)$$

where N is the total number of predictions, $\mathrm{acc}(B_m)$ is the average accuracy of predictions in bin B_m, and $\mathrm{conf}(B_m)$ is the average predicted confidence.

Prompt used for UnbiasedVQA QA generation

```
System prompt:
You are a helpful assistant that generates
    multiple choice questions and
    corresponding groundtruth answers
    based on image descriptions and the
    image.

User prompt:
Based on this image and its description
    below, generate 3 different multiple
    choice questions with 10 answer
    choices each. Each question can
    address different aspects of the image
    and must remain relevant to the
    provided content. The answer choices
    should be correct, but only one should
    be correct.

The question should cover a wide range of
    aspects of the image, including but
    not limited to:
1. Overall Content: Summarize the main
    focus or theme of the image.
2. Human Description: Describe each
    person's appearance, clothing,
    emotional expressions, gender,
    approximate age, physique, skin/hair
    color, accessories, and other
    distinguishing features.
3. Object Description: Identify and
    characterize visible objects, noting
    their attributes (color, texture,
    function).
4. Scene Description: Depict the broader
    setting (indoor/outdoor, location
    type, time of day, season, weather,
    lighting).
5. Events and Interactions: Highlight all
    significant events, interactions
    between participants, how objects are
    used by participants and any other
    distinguishing events.
6. Spatial Order: Clarify how people and
    objects are positioned relative to
    each other.
7. Count: Provide the number of people,
    objects, and notable visual elements.
8. Camera Direction: Mention notable camera
    angles, perspectives, or framing
    styles.

Below is the image description:
{description}
```

D z-test experiments

Table 6 compares train–test accuracy gaps across COCO-based benchmarks and the proposed UnbiasedVQA benchmark. Several MLLMs exhibit statistically significant train–test discrepancies on A-OKVQA and Visual7W, suggesting stronger dataset-specific bias in COCO-based benchmarks.

Table 6: **Train–test accuracy comparison across COCO-based benchmarks and the proposed UnbiasedVQA benchmark:** A-OKVQA and Visual7W show statistically significant train–test gaps for several MLLMs, suggesting dataset bias in COCO-based benchmarks. In contrast, UnbiasedVQA shows no statistically significant train–test difference across evaluated models, indicating a more reliable benchmark for assessing general scene understanding.

Dataset	Train/Test Size	Model	Size	Train Acc.	Test Acc.	Train Acc. − Test Acc.	p-value
A-OKVQA	17k / 1.1k	LLaVA-1.5	13B	86.15	80.00	6.15	$\mathbf{8.1 \times 10^{-9}}$
		Phi-4	5.6B	96.12	82.71	13.41	$\mathbf{< 10^{-80}}$
		Qwen2.5VL	7B	87.44	86.72	0.72	0.48
		InternVL3	8B	89.39	84.28	5.11	$\mathbf{8.2 \times 10^{-8}}$
		GPT-4o	–	86.19	86.81	−0.62	0.55
Visual7W	70k / 28k	LLaVA-1.6	7B	81.44	81.10	0.34	0.21
		Phi-4	5.6B	84.94	83.76	1.18	$\mathbf{3.5 \times 10^{-6}}$
		Qwen2.5VL	7B	88.04	87.20	0.84	$\mathbf{2.6 \times 10^{-4}}$
		InternVL3	8B	90.30	88.72	1.58	$\mathbf{1.7 \times 10^{-13}}$
		PLM	8B	87.64	86.20	1.44	$\mathbf{1.26 \times 10^{-9}}$
UnbiasedVQA (Proposed)	23.9k / 12.3k	LLaVA-1.6	7B	72.31	72.92	−0.61	0.22
		Phi-4	5.6B	80.55	80.61	−0.06	0.67
		Qwen2.5VL	7B	84.25	84.86	−0.61	0.13
		InternVL3	8B	82.12	82.76	−0.64	0.13
		PLM	8B	83.24	83.56	−0.32	0.44

In contrast, UnbiasedVQA maintains consistent train/test performance across evaluated models, with no statistically significant train–test differences observed. Note that although the absolute train–test accuracy gaps for Visual7W are relatively smaller than those for A-OKVQA, the corresponding p-values remain very small due to the substantially larger sample size used in the z-test for Visual7W.

(1) Question: How is the cat positioned on the couch?

Choices:
0. Sleeping on its back
1. Eating food
2. Climbing the couch
3. Running across the couch
4. Sitting upright
5. Sitting with its tail straight
6. Jumping off the couch
7. Playing with a toy
8. Standing on all fours
9. Laying down with its body curled

(2) Question: What objects are placed in the center hole of the tractor tire?

Choices:
0. Two rugby balls and one tennis ball
1. Two volleyballs and one basketball
2. Three tennis balls
3. Two cricket balls and one rugby bal
4. Two basketballs and one volleyball
5. Three hockey pucks
6. Three golf balls
7. Three baseballs
8. Two soccer balls and one football
9. Two footballs and one soccer ball

(3) Question: What flags are visible on the steamboat model?

Choices:
0. South African and Egyptian flags
1. Dutch and Belgian flags
2. Canadian and Mexican flags
3. Russian and Indian flags
4. Brazilian and Argentinian flags
5. Australian and New Zealand flags
6. German and Italian flags
7. British and French flags
8. American and Spanish flags
9. Japanese and Chinese flags

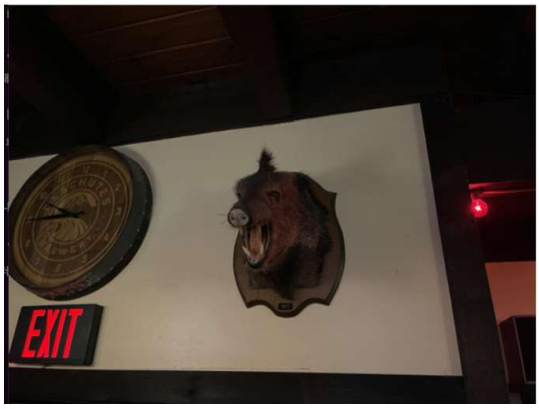
(4) Question: What shape is the top sign on the post?

Choices:
0. Square
1. Pentagon
2. Star
3. Rectangle
4. Diamond
5. Circle
6. Octagon
7. Triangle
8. Hexagon
9. Oval

Figure 7: **Examples of validated QA pairs included in UnbiasedVQA:** These samples illustrate description-grounded questions that capture scene-level understanding, such as object attributes, spatial relations, and overall context, with a single unambiguous correct answer.

(1) Question: What is the sculpture using to play the drum?

Choices:
0. Its right leg
1. Its right hand ← GT
2. Its left antenna
3. Its left foot
4. Its left hand
5. Its right eye
6. Its right antenna
7. Its left leg
8. Its right foot
9. Its left eye



(2) Question: What time is displayed on the clock in the image?

Choices:
0. 6:30
1. 3:15
2. 12:00
3. 10:45 ← GT
4. 4:20
5. 7:50
6. 11:30
7. 2:25
8. 1:00
9. 9:45

(3) Question: How many large objects are laying on the dirt area beside the lawn?

Choices:
0. One
1. Ten
2. Five
3. Nine
4. Seven
5. Six
6. Three ← GT
7. Two
8. Four
9. Eight

(4) Question: What is silhouetted against the sky in the background on the right?

Choices:
0. A satellite dish
1. A skyscraper
2. A helicopter
3. A mountain
4. A wind turbine
5. A radio tower
6. A big crane ← GT
7. A hot air ballon
8. A tall tree
9. A billboard

Figure 8: **Examples of filtered QA pairs during validation in UnbiasedVQA:** (1) Perspective ambiguity (2) Incorrect ground truth (3) Vague question (4) multiple plausible answers.