Critical Analysis: MRAG-BENCH: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models

Kai-Yu Lu

1 Methodological Strengths

1.1 Vision-centric benchmark design

MRAG-BENCH introduces a retrieval-augmented evaluation benchmark that is explicitly vision-centric. The benchmark targets scenarios in which external *visual* knowledge is more useful than textual knowledge, which fills a clear gap in existing retrieval-augmented generation (RAG) benchmarks.

The benchmark focuses on large vision-language models (LVLMs), that is, models jointly processing images and text at scale, and evaluates their ability to exploit retrieved images for multiple-choice question answering (MCQA). The design is structured around nine carefully defined multimodal RAG scenarios, grouped into two major aspects:

- **Perspective aspect** (57.5% of questions):
  - **ANGLE** (23.8%): different viewpoints of the same object.
  - **PARTIAL** (18.2%): only partial views are visible.
  - **SCOPE** (7.5%): different zoom levels or crop scopes.
  - **OCCLUSION** (8.0%): objects are partially occluded.
- **Transformative aspect** (33.6% of questions):
  - **TEMPORAL** (11.0%): objects across time, such as under construction vs finished.
  - **DEFORMATION** (7.5%): deformed or distorted objects.
  - **INCOMPLETE** (7.5%): missing parts, such as missing keys on a keyboard.
  - **BIOLOGICAL** (7.5%): biological transformations, such as oxidation or decay.
- **OTHERS** (8.9%): primarily geographic origin reasoning using region-specific visual cues.

The explicit decomposition into perspective and transformative dimensions provides a conceptually clear framework for analyzing failure modes related to viewpoint variation and physical transformation. This structure goes beyond generic question-answering benchmarks and enables targeted diagnosis of LVLM limitations.

1.2 Systematic data construction and curation

The benchmark construction follows well-defined data sourcing and quality control procedures.

Image sources. For the perspective scenarios, MRAG-BENCH relies on established object recognition datasets: ImageNet, Oxford Flowers102, and StanfordCars. These datasets provide diverse, high-quality category-level images. For the transformative scenarios, images are scraped from the web using Bing Image Search, and for geographic scenarios the benchmark utilizes the GeoDE dataset, which is explicitly designed for geographically diverse object recognition.

Ground-truth corpus and query images. For each scenario, there is a distinction between:

- A **ground-truth image corpus** of representative, informative images that encode canonical visual knowledge for each object or state.
- **Query images** that deliberately challenge LVLMs by using less typical viewpoints, occlusions, deformations, or temporal states.

Human annotators manually select ground-truth images to ensure that they are visually informative and align with human expectations for each category. For many classes, five representative ground-truth images are selected. For **PARTIAL** and **INCOMPLETE** scenarios, query images are generated by controlled cropping and then filtered by human inspection to ensure that the target object is still clearly identifiable and that the crop reflects the intended scenario definition.

The benchmark contains 16,130 images and 1,353 human-annotated multiple-choice questions, each with on average four answer choices. All questions are newly annotated. This scale is moderate but sufficient for controlled, scenario-level analysis, rather than merely large-scale leaderboard ranking.

Quality control. The construction pipeline incorporates both automatic and manual quality checks:

- Automatic checks ensure valid MCQA format, image validity, and removal of redundant images.
- Manual checks by domain experts verify the alignment between query images, ground-truth image sets, and the question content, and filter ambiguous or misaligned items.

This combination of automatic and manual validation provides a relatively high degree of reliability and consistency, which is essential for fine-grained analysis across scenarios.

1.3 Well-specified evaluation protocol

The evaluation protocol is clearly defined and reproducible. The retrieval-augmented pipeline is formulated as follows: given a query tuple $Q = (\text{query image}, \text{text question})$, a multimodal retriever R returns a set of images $I = [i_1, \ldots, i_N]$, and an LVLM M produces an answer based on (Q, I). The use of multiple-choice questions simplifies evaluation to accuracy, avoiding ambiguities of free-form generative evaluation.

Three conditions are systematically compared:

- **No RAG**: model answers using only the query image and question.
- **Retrieved RAG**: model answers using retrieved image examples.
- **Ground-truth (GT) RAG**: model answers using human-selected ground-truth image examples.

By using the same CLIP-based retriever across LVLMs for the main table and fixing the number of retrieved images (five for most scenarios, one for the incomplete scenario), the study carefully controls for retrieval variability and isolates the effect of visual augmentation on different model architectures.

1.4 Broad model coverage and human baseline

The experimental evaluation covers 14 LVLMs:

- Four proprietary models: GPT-4o, GPT-4-Turbo, Gemini Pro, and Claude 3.5 Sonnet.
- Ten open-source models, including LLaVA-OneVision, LLaVA-Next-Interleave, mPLUG-Owl3, VILA, Deepseek-VL, and Mantis variants.

This coverage allows a direct comparison between strong proprietary systems and current open-source LVLMs under exactly the same retrieval and MCQA conditions.

In addition, MRAG-BENCH includes a human baseline. Without any retrieved images, human subjects obtain 38.47% accuracy. With retrieved images from the benchmark corpus, human performance rises to 61.38%, and with ground-truth image examples, human performance reaches 71.63%. These numbers are directly comparable to LVLM results and provide a realistic reference for how much visual retrieval can help a human reasoner.

1.5 Rich analysis of retrieval configurations

Beyond the main accuracy table, the paper performs several focused analyses:

- **Image vs text retrieval:** For selected models, the study compares performance when augmenting with retrieved images versus retrieved text. On MRAG-BENCH, image retrieval yields larger gains than text retrieval, especially under ground-truth augmentation. This empirically validates the central claim that MRAG-BENCH is vision-centric rather than text-centric.
- **Retriever comparison and position bias:** The study evaluates four state-of-the-art multimodal retrievers (CLIP, MagicLens, E5-V, VISTA) and reports a strong 95% correlation between retriever Recall@5 and downstream LVLM accuracy for LLaVA-Next-Interleave. This is a methodologically strong connection between retrieval quality and question-answering performance.
- **Number of ground-truth images:** The analysis varies the number of ground-truth image examples and shows that performance increases with more images, peaking around ten images and saturating afterward. This gives a concrete empirical view of how context length and image count interact.

These analyses move beyond simple single-setting evaluation and provide insight into how RAG configurations affect performance.

2 Key Limitations

2.1 Scale and diversity of the benchmark

Although MRAG-BENCH covers nine scenarios and 1,353 questions, the absolute scale is relatively modest compared with many modern multimodal benchmarks. Several limitations follow:

- The number of unique questions (375) and unique answers (663) suggests that many questions share structural similarity. This may restrict the diversity of reasoning patterns that can be evaluated.
- Most scenarios focus on object recognition, viewpoint variation, and physical transformations of single objects. High-level compositional reasoning, complex multi-object interactions, and long-horizon tasks are less represented.

- All questions follow a multiple-choice format with short answers and average question length around eight tokens. This design simplifies evaluation but does not stress models on open-ended explanation, justification, or multi-step reasoning.

Consequently, although MRAG-BENCH is well-suited for diagnosing visual retrieval and recognition robustness, it provides limited coverage of more complex multimodal reasoning skills.

2.2 Domain bias in data sources

The benchmark draws heavily from specific datasets and web sources:

- **Perspective scenarios** rely on ImageNet, Oxford Flowers102, and StanfordCars. These datasets primarily represent curated, centered objects with relatively clean backgrounds and can under-represent challenging real-world clutter or long-tailed categories.
- **Transformative scenarios** are collected by web scraping using predefined keywords. This process may introduce bias toward popular visual patterns and well-photographed objects, while undersampling rare transformations or harder edge cases.
- **Geographic scenarios** rely on GeoDE, which is designed for geographic diversity but still reflects particular choices of categories and regions.

As a result, MRAG-BENCH is more representative of object-centric, image-classification-style domains rather than unconstrained real-world environments or application-specific domains such as medical images or industrial inspection.

2.3 Limited exploration of textual RAG and hybrid RAG

The benchmark is intentionally vision-centric, and the text-based RAG baselines are fairly limited. The main text-centric comparison is performed on a subset of models and settings, and the design does not deeply explore:

- Retrieval from large textual corpora such as Wikipedia at scale.
- Hybrid combinations of text and images where both modalities jointly contribute complementary evidence.
- The effect of noisy textual retrieval or contradictions between text and images.

This leaves open the question of how MRAG-BENCH scenarios would behave under realistic large-scale multimodal RAG systems that flexibly mix text and visual evidence.

2.4 Evaluation metrics and error analysis

The primary metric is accuracy on MCQA, which, while clear and easy to interpret, has several limitations:

- Accuracy does not differentiate between near-miss choices and completely incorrect answers, nor does it reflect model confidence or calibration.
- The paper provides limited fine-grained error categorization beyond scenario-level accuracy. Detailed qualitative error analysis is only briefly illustrated through example cases rather than systematically quantified.

- There is no explicit computational cost analysis, even though the use of multiple images and complex LVLMs introduces significant compute requirements.

More nuanced metrics such as calibration error, robustness under input perturbations, or retrieval sensitivity are not reported, which constrains the interpretability of model behavior.

2.5 Restricted analysis of training-time factors

MRAG-BENCH is purely an evaluation benchmark. The paper does not train models on MRAG-BENCH, nor does it systematically vary training recipes. As a result, the study does not disentangle how much of the observed performance gaps arise from:

- Architectural limitations in handling multi-image inputs.
- The underlying pretraining corpora and image distributions.
- Fine-tuning strategies for multimodal RAG.

This limits causal interpretation of why proprietary models outperform open-source models in leveraging retrieved images.

3 Technical Bottlenecks

3.1 Limited utilization of retrieved visual knowledge

A central finding is that even the strongest LVLMs exhibit limited gains from retrieved images. GPT-4o achieves 68.68% accuracy without RAG, 68.96% with retrieved images, and 74.50% with ground-truth images. Thus, retrieved images provide only a 0.28% absolute gain, and even perfect ground-truth examples yield a modest 5.82% improvement.

In contrast, humans improve by 22.91% with retrieved images and 33.16% with ground-truth images. This indicates a fundamental bottleneck: current LVLMs do not exploit additional visual context nearly as effectively as humans. Possible technical causes include:

- Insufficient architecture capacity or inductive bias to compare and align multiple images.
- Shallow integration of retrieved images, treating them as additional tokens rather than as structured visual evidence.
- Over-reliance on parametric knowledge, with models defaulting to internal priors rather than carefully examining retrieved images.

3.2 Fragility under noisy or imperfect retrieval

Many open-source LVLMs degrade when using retrieved images. For example, models such as VILA, Deepseek-VL, Mantis, and mPLUG-Owl3 often exhibit negative accuracy deltas under retrieved RAG compared with no RAG, despite clear gains under ground-truth RAG.

This exposes a strong sensitivity to retrieval noise and irrelevance. The LVLMs tend to:

- Over-trust retrieved images even when they are off-topic or misleading.
- Fail to filter or down-weight low-quality examples.

- Struggle with conflicts between retrieved images and the query image.

Such behavior suggests that current LVLMs lack robust mechanisms for reliability-aware evidence selection and for distinguishing high-quality from low-quality retrievals.

3.3 Dependence on retriever quality and position bias

The 95% correlation between retriever Recall@5 and downstream accuracy for LLaVA-Next-Interleave demonstrates that retrieval quality strongly controls performance. Additionally, the study observes non-trivial gaps between retrievers with similar Recall@5 but different ordering of retrieved images. This is described as position bias: models are sensitive to the order in which images are presented and may over-focus on early examples.

This reveals two bottlenecks:

- LVLMs do not yet robustly aggregate information across a set of unordered images, which would ideally be permutation invariant or permutation robust.
- Current retrieval scoring and ranking are not explicitly optimized for downstream reasoning, only for similarity search.

3.4 Context length and scaling with number of images

The analysis on the number of ground-truth examples shows that LLaVA-Next-Interleave gains from one image up to about ten images, after which performance saturates. The improvement with one ground-truth example is already substantial (around 5.64%), and the difference between ten and twenty images is very small (around 0.29%).

This pattern indicates:

- LVLMs have limited ability to exploit very large visual contexts; there is a saturation point beyond which additional images do not help and may even create distraction.
- The architecture and attention mechanisms are not specifically designed to scale gracefully with many images, leading to diminishing returns.

Thus, handling large retrieved image sets in a principled, scalable manner remains a technical challenge.

3.5 Scenario-specific weaknesses

The scenario-level results reveal systematic weaknesses:

- Transformative scenarios such as TEMPORAL, DEFORMATION, and BIOLOGICAL remain challenging even for strong proprietary models, especially under noisy retrieval.
- The INCOMPLETE scenario introduces structural reasoning about missing parts and layouts, where models show only modest gains.
- The OTHERS scenario, which involves geographic origin reasoning, exposes dependence on geographically balanced training data and retrieval corpora.

These patterns suggest that LVLMs are still far from robustly modeling real-world variations such as aging, damage, composition changes, and geographic styles.

4 Research Implications

4.1 Reframing multimodal RAG evaluation

MRAG-BENCH shifts attention from text-centric retrieval to vision-centric retrieval in multimodal RAG. The empirical evidence that visual retrieval is more helpful than text retrieval in this benchmark highlights the importance of designing RAG systems that can flexibly choose the appropriate modality of retrieval depending on the task.

The benchmark demonstrates that:

- LVLMs are not yet reliable visual evidence aggregators, especially under retrieval noise.
- Properly structured vision-centric benchmarks can reveal weaknesses that standard text-based or single-image benchmarks do not expose.

4.2 Understanding the gap between proprietary and open-source LVLMs

Across MRAG-BENCH, proprietary models such as GPT-4o and Gemini Pro substantially outperform open-source models. The performance gap is especially pronounced under noisy retrieved RAG, where proprietary models still obtain positive or small improvements, while many open-source models deteriorate.

This suggests that proprietary models may incorporate:

- Stronger parametric visual knowledge.
- More robust mechanisms for cross-image comparison and evidence selection.
- Better training regimes on multi-image and retrieval-like tasks.

The benchmark thus provides a clear empirical target for open-source LVLM development.

4.3 Highlighting the under-explored dimension of visual position bias

The observed position bias in retrieved images parallels findings in text-only RAG but in the visual domain. This has several implications:

- Evaluations that do not randomize or stress-test retrieval order may overestimate robustness.
- Multi-image encoders and attention mechanisms should be designed or trained to be less sensitive to ordering.
- Retrieval pipelines may need explicit mechanisms to diversify the order and presentation of images to mitigate bias.

4.4 Evidence for calibration and reliance issues

The relatively small improvement for GPT-4o under ground-truth RAG, compared with the large human improvement, suggests that LVLMs are under-using external evidence. At the same time, the degradation of many open-source models under noisy RAG suggests over-reliance on retrieved content without sufficient skepticism.

These two phenomena indicate miscalibration in how LVLMs weigh internal parametric knowledge versus external retrieved evidence, which has direct implications for deploying RAG systems in safety-critical settings.

5 Potential Research Directions

5.1 Architectures for multi-image reasoning

One natural direction is to design LVLM architectures that explicitly support:

- Permutation-invariant or permutation-robust image set encoders.
- Hierarchical attention mechanisms that first reason within each image and then across images.
- Memory or graph-style representations that track correspondences and transformations between query images and retrieved examples.

Such architectures may reduce position bias and improve utilization of additional visual context.

5.2 Retrieval-aware training and joint optimization

The strong dependence on retriever quality suggests that retrievers and LVLMs should be trained jointly or in a tightly coupled fashion. Research directions include:

- Joint training of retrievers and LVLMs using MRAG-BENCH-style supervision, where retrieval is directly optimized for downstream MCQA accuracy.
- Differentiable retrieval objectives that propagate gradients from the answer loss back into retrieval scoring.
- Curriculum learning that gradually increases retrieval difficulty and noise.

5.3 Robustness to retrieval noise and evidence selection

Mitigating the degradation under noisy retrieved images is critical. Potential directions include:

- Reliability-aware attention mechanisms that estimate the usefulness of each retrieved image and down-weight irrelevant or conflicting ones.
- Auxiliary training objectives that penalize over-reliance on misleading images, for instance by adding adversarial or counterfactual retrievals.
- Confidence calibration methods that align model confidence with the quality and consistency of supporting visual evidence.

5.4 Adaptive control of the number of retrieved images

The analysis on the number of ground-truth images indicates that not all questions benefit from many images. Future work can explore:

- Policies that dynamically decide how many images to retrieve based on question complexity and intermediate uncertainty estimates.
- Multi-stage retrieval, in which a small initial set of images is used to refine the query and selectively retrieve more relevant images.
- Compression or summarization of multiple images into compact latent representations, which can be used for efficient downstream reasoning.

5.5 Extending MRAG-BENCH to richer tasks and domains

The benchmark naturally invites extensions:

- Open-ended question answering, explanation generation, and step-by-step reasoning grounded in retrieved images.
- Expansion to domain-specific scenarios, such as medical, industrial, or scientific imaging, where transformations and temporal dynamics are critical.
- Inclusion of video and temporal sequences, linking the TEMPORAL and DEFORMATION scenarios to genuine time-series visual data.

Such extensions would further stress-test LVLMs on real-world deployment requirements.

6 Conclusion

MRAG-BENCH provides a carefully designed, vision-centric benchmark for evaluating retrieval-augmented large vision-language models. Its strengths lie in the explicit scenario structure, systematic data collection, controlled RAG evaluation protocol, and rich analyses relating retriever quality, number of images, and LVLM performance.

At the same time, the benchmark reveals several important limitations and bottlenecks. LVLMs currently under-utilize retrieved visual knowledge, are fragile under retrieval noise, exhibit position bias, and struggle with transformative and incomplete visual scenarios. These findings highlight substantial gaps between current LVLM capabilities and human-level use of visual context.

Overall, MRAG-BENCH serves both as a diagnostic tool and as a roadmap for future research. It underscores the need for multi-image reasoning architectures, retrieval-aware training, robustness to noisy evidence, and adaptive retrieval strategies. Addressing these challenges will be essential for building reliable, vision-centric retrieval-augmented multimodal systems.