# Technical Summary:
# MRAG-BENCH: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models

**Kai-Yu Lu**

## 1 Research Problem and Motivation

### 1.1 Research Problem

Retrieval-augmented generation is a widely used paradigm for enhancing large language models and large vision language models, by allowing these models to access external knowledge during inference. Existing multimodal retrieval benchmarks largely evaluate whether models can retrieve and exploit external *textual* knowledge for visual question answering. In these settings, a model receives a single image, a textual question, and a set of retrieved text passages, and must answer the question based on this combined information.

However, many real-world situations are *vision-centric*. In such situations, retrieving additional images is more beneficial or more accessible than retrieving textual documents. The central research problem addressed by MRAG-BENCH is the following: how well can large vision language models utilize *retrieved visual knowledge*, rather than only retrieved text, in retrieval-augmented multimodal reasoning.

### 1.2 Motivation and Research Gap

Previous benchmarks, including OK-VQA, A-OKVQA, WebQA, Encyclopedia VQA, and InfoSeek, are primarily text-centric. These benchmarks focus on questions that cannot be answered from the image alone and require external world knowledge retrieved from text sources such as Wikipedia or curated web corpora. The external knowledge modality is predominantly text, and models are evaluated on their ability to integrate text with visual context.

This text-centric focus leads to two important gaps.

- First, these benchmarks do not target scenarios where additional images are intrinsically more informative than additional text. For example, recognizing a specific car model from an uncommon top-down view is difficult even with detailed textual descriptions, because such descriptions rarely capture the exact appearance from that viewpoint. In contrast, retrieving multiple images of the same car from common viewpoints makes recognition considerably easier.
- Second, these benchmarks do not systematically categorize challenging visual conditions such as rare viewpoints, occlusion, deformation, incomplete structures, or biological transformations. As a result, they cannot reveal how well large vision language models leverage additional images in diverse vision-centric scenarios.

Real-world applications of multimodal systems frequently involve visual queries that are best resolved by examining additional images. For example, understanding how fruit decays, how buildings look before and after construction, or how objects appear when partially occluded are tasks that rely on visual exemplars. Textual descriptions are often insufficient to capture subtle visual patterns in these cases.

The research gap is therefore the lack of a systematic, vision-centric retrieval-augmented benchmark that

- explicitly targets scenarios where retrieved visual knowledge is more useful than textual knowledge, and
- evaluates how much large vision language models and humans actually benefit from additional images in these scenarios.

### 1.3 Research Objectives

The MRAG-BENCH study has the following objectives.

- To construct a benchmark that emphasizes vision-centric retrieval-augmented reasoning, where additional images are the primary external knowledge source.
- To define and instantiate a taxonomy of nine scenarios, covering perspective changes and transformative appearance changes, and to provide high-quality image corpora and multiple-choice questions for these scenarios.
- To systematically evaluate a broad set of large vision language models and human participants under multiple retrieval configurations, in order to quantify the benefits of visual versus textual knowledge and to analyze the effect of retriever quality and context size.

## 2 Related Work

### 2.1 Multimodal Retrieval-Augmented Benchmarks

Several benchmarks have been developed for retrieval-augmented visual question answering. OK-VQA focuses on questions that require external world knowledge beyond the image. A-OKVQA extends OK-VQA with additional knowledge types. MultiModalQA, ManyModalQA, WebQA, Encyclopedia VQA, and InfoSeek rely heavily on encyclopedic or web-scale textual corpora. In these benchmarks, knowledge is predominantly textual, and retrieval operates over text passages or captions.

Table 1 summarizes the contrast. Existing benchmarks mainly use text as the external knowledge modality, rely on Wikipedia or other web sources, and do not provide explicit multi-image input or scenario-level categorization. MRAG-BENCH instead uses images as the primary retrieved modality, supports multi-image inputs, and defines nine scenarios to cover diverse vision-centric conditions.

### 2.2 Large Vision Language Models

Large vision language models such as Flamingo, Emu, Idefics, VILA, LLaVA variants, Mantis, mPLUG-Owl, Deepseek-VL, Pixtral, GPT-4 variants, Gemini Pro, and Claude 3.5 Sonnet demonstrate strong performance on a wide range of multimodal tasks. Many of these models are trained on interleaved image-text corpora and can process multiple images jointly. Instruction tuning further improves their ability to follow natural language instructions while reasoning over visual inputs.

Existing evaluations typically focus on single-image tasks, text-centric retrieval, or generic multimodal benchmarks. The ability of large vision language models to use multiple retrieved images as external knowledge in a controlled, vision-centric benchmark has not been systematically evaluated before MRAG-BENCH.

| Benchmark | Knowledge modality | Knowledge source | Multi-image input | Scenario categorization |
|---|---|---|---|---|
| K-VQA | Text | Wikipedia | No | No |
| OK-VQA | Text | Wikipedia | No | No |
| MultiModalQA | Text | Wikipedia | No | No |
| ManyModalQA | Text | Wikipedia | No | Yes |
| A-OKVQA | Text | World knowledge | No | No |
| ViQuAE | Text | Wikipedia | No | No |
| WebQA | Text and caption | Wikipedia and web | No | No |
| Encyclopedia VQA | Text | Wikipedia | No | No |
| InfoSeek | Text | Wikipedia | No | No |
| MRAG-BENCH | Image | Image corpora and datasets | Yes | Yes (nine scenarios) |

Table 1: Comparison between MRAG-BENCH and representative prior retrieval-augmented benchmarks.

### 2.3 Retrieval-Augmented Multimodal Models

Retrieval-augmented generation was originally introduced for language models to overcome limitations of parametric memory by incorporating external document retrieval during generation. Subsequent work extended this idea to multimodal retrieval, where both images and text can serve as external knowledge. Recent systems retrieve web images or combined image-text documents to support question answering and reasoning.

Despite these advances, prior work does not provide a dedicated, vision-centric benchmark that isolates cases where images are more helpful than text, nor does it compare human and model performance in such scenarios. MRAG-BENCH is designed to fill this gap.

## 3 Dataset Construction

### 3.1 Design Principles and Global Statistics

MRAG-BENCH is constructed to evaluate vision-centric retrieval-augmented reasoning in real-world scenarios. The dataset construction follows four principles.

- The benchmark focuses on scenarios where visually augmented information is genuinely useful.
- The benchmark covers diverse types of visual reasoning, including viewpoint variation, physical deformation, incompleteness, and biological transformations.
- For each question, a set of high-quality ground-truth images is provided to approximate human visual knowledge.
- The benchmark supports robust and deterministic evaluation of large vision language models.

MRAG-BENCH includes 16,130 images and 1,353 human-annotated multiple-choice questions across nine scenarios. All questions are newly annotated. Key statistics are summarized in Table 2.

Questions are divided into three high-level types. Perspective scenarios account for 57.5% of the questions, transformative scenarios account for 33.6%, and others account for 8.9%. Within these types, nine specific scenarios are defined, each comprising between 7.5% and 23.8% of the dataset.

| Statistic | Value |
|---|---|
| Total questions | 1,353 |
| Question format | Multiple-choice visual question answering |
| Number of scenarios | 9 |
| Total images | 16,130 |
| Unique images | 16,130 |
| Human-selected ground-truth images | 9,673 |
| Average image size (pixels) | 1076 × 851 |
| Maximum question length (tokens) | 20 |
| Maximum answer length (tokens) | 9 |
| Average question length (tokens) | 8.03 |
| Average answer length (tokens) | 2.16 |
| Average number of answer choices | 4 |

Table 2: Core statistics of MRAG-BENCH.

## 3.2 Scenario Taxonomy

### 3.2.1 Perspective Scenarios

Perspective scenarios concern variation in viewpoint, visibility, and resolution of visual entities.

- **ANGLE** (23.8%): evaluates the ability to use knowledge of common shooting angles to identify and reason about long-tailed, less common viewpoints of objects.
- **PARTIAL** (18.2%): evaluates recognition when only a partial view of the object is visible, using complete exemplars as visual knowledge.
- **SCOPE** (7.5%): evaluates recognition in low-resolution or long-range views, where the target is small and high-resolution close-ups serve as helpful exemplars.
- **OCCLUSION** (8.0%): evaluates reasoning when objects are occluded or partially hidden, and unobstructed ground-truth images provide critical cues.

### 3.2.2 Transformative Scenarios

Transformative scenarios involve physical or temporal changes to an object's appearance.

- **TEMPORAL** (11.0%): focuses on temporal changes, such as buildings under construction versus completed buildings, or animals at different life stages.
- **DEFORMATION** (7.5%): covers deformed objects, such as damaged vehicles, where intact ground-truth images support recognition.
- **INCOMPLETE** (7.5%): focuses on incomplete objects with missing parts, such as keyboards with absent keys, and requires layout comparison.
- **BIOLOGICAL** (7.5%): addresses biological transformations such as mold growth and oxidation, and asks about likely or unlikely visual outcomes.

### 3.2.3 Others Scenario

The **OTHERS** scenario (8.9%) evaluates the use of geographic image knowledge. For example, given an artifact and candidate regions, the model must infer the correct region by comparing the query with ground-truth images from different regions.

## 3.3 Image Sources and Collection Pipeline

### 3.3.1 Perspective Aspect

Perspective scenarios draw images from three main public datasets.

- **ImageNet**: a large-scale object classification dataset used to provide diverse categories.
- **Oxford Flowers102**: a dataset of 102 flower categories with fine-grained labels.
- **StanfordCars**: a fine-grained dataset of car makes and models.

For each selected class, validation images are examined and images that do not provide sufficient visual information for recognition are removed. From the remaining candidates, annotators select five representative images per class as ground-truth exemplars that cover diverse appearances.

For the **ANGLE**, **SCOPE**, and **OCCLUSION** scenarios, query images are manually selected to match the scenario definitions. For the **PARTIAL** scenario, images are randomly cropped by 50% in both height and width. Human annotators then filter out crops that do not contain a dominant view of the target object. This process is repeated until suitable partial views are obtained. On average, perspective scenarios provide approximately 20.4 ground-truth images per question.

### 3.3.2 Transformative Aspect

Transformative scenarios use web-scraped images.

- Images are collected with Bing Image Search using predefined keywords that describe objects under specific transformations, such as damaged cars or decayed fruit.
- Candidate keywords are retained when they yield clear pairs of query and ground-truth images illustrating the intended transformation, resulting in 74% of the keyword names being retained.
- For each retained keyword, up to 15 images are downloaded. Human annotators filter out low-quality or off-topic images.
- For each question, an average of 5.9 images is retained and five are selected as ground-truth exemplars.

### 3.3.3 Others Scenario from GeoDE

The **OTHERS** scenario uses the GeoDE dataset, in which objects are labeled with geographic regions. For each object category, three regions out of six available regions are sampled as answer options, and an image from one of these regions is chosen as the query image. Ground-truth exemplars for the correct region serve as the image knowledge for retrieval.

| Name | Modality | Role in MRAG-BENCH | Characteristics |
|---|---|---|---|
| ImageNet | Image | Perspective scenarios | Large-scale object classification |
| Oxford Flowers102 | Image | Perspective scenarios | Fine-grained flower categories |
| StanfordCars | Image | Perspective and deformation scenarios | Fine-grained car models |
| GeoDE | Image | Others scenario | Objects labeled by geographic region |
| Bing Image Search | Image | Transformative scenarios | Web images under transformations |
| Wikipedia dump (2023-07-01) | Text | Textual retrieval baseline | Encyclopedic text corpus |

Table 3: External datasets and corpora used in MRAG-BENCH.

### 3.4 Quality Control

Quality control is performed in two stages.

- An automatic check verifies that each instance follows the multiple-choice format, that all images are valid, and that redundant images are removed from the corpus.
- A manual examination by two experts inspects each instance to ensure consistency between query images and ground-truth exemplars, and to revise or discard ambiguous or misaligned questions.

### 3.5 External Datasets and Knowledge Sources

Table 3 summarizes external datasets and corpora used by MRAG-BENCH.

## 4 Query Protocol and Task Definitions

### 4.1 Task Format

Each MRAG-BENCH instance is a multiple-choice visual question answering task. An instance consists of

- a query image that illustrates the visual scenario,
- a textual question about the image, and
- a set of answer options, with exactly one correct answer.

The goal of a model or human participant is to select the correct answer option.

### 4.2 Retrieval-Augmented Query Protocol

The retrieval-augmented query protocol is formulated in terms of a query tuple and a multimodal retriever.

Let $Q$ denote the query, which consists of a query image and a textual question. A multimodal retriever $R$ maps $Q$ to a list of retrieved images $I$ as follows

$$I = R(Q) = [i_1, i_2, \ldots, i_N], \tag{1}$$

where $I$ is the ordered list of retrieved images, $i_k$ is the $k$-th retrieved image, and $N$ is the number of retrieved images, typically five. This equation states that the retriever processes the combined query and returns the top $N$ images from the image corpus that are most relevant according to its similarity function.

A large vision language model $M$ then receives the query and the retrieved images and produces a predicted answer $a$ among the multiple-choice options

$$a = M(Q, I). \tag{2}$$

In this equation, $M$ denotes the parameterized multimodal model, $Q$ is the original query (image and text), $I$ is the list of retrieved images from Equation (1), and $a$ is the final discrete answer. Intuitively, the model takes the question, the query image, and the retrieved images as joint input and outputs the choice that best matches all available visual and textual evidence.

## 4.3 Evaluation Metrics

### 4.3.1 Accuracy

The primary metric for question answering is accuracy

$$\mathrm{Acc} = \frac{\text{number of correctly answered questions}}{\text{total number of questions}}. \tag{3}$$

In this formula, the numerator counts how many questions the model or human answers correctly, and the denominator is the total number of questions in the evaluation set. The accuracy value can be interpreted as a test score that reflects the fraction of correct decisions.

### 4.3.2 Retriever Recall@5

Retriever quality is analyzed using Recall@5. For each question, the retriever returns the top five images. Recall@5 is defined as

$$\mathrm{Recall@5} = \frac{\text{number of questions with at least one ground-truth image among top five}}{\text{total number of questions}}. \tag{4}$$

In this formula, the numerator counts how many questions have at least one human-selected ground-truth exemplar within the top five retrieved images, and the denominator is again the total number of questions. A higher Recall@5 value indicates that the retriever more frequently places relevant images within its highest-ranked candidates, which is essential for effective retrieval-augmented reasoning.

## 4.4 Evaluation Settings

MRAG-BENCH supports several evaluation configurations.

- **No RAG**: The model only receives the query image and the textual question without any retrieved images or text.
- **Retrieved Image RAG**: CLIP-based multimodal retrieval is used to fetch the top five images, which are given to the model together with the query.
- **Ground-Truth Image RAG**: Five curated human-selected ground-truth images are provided for each question, serving as ideal image knowledge.
- **Retrieved Text RAG**: Text passages are retrieved from a Wikipedia dump using the same query and appended to the input.

| Model | Type | Approximate scale | Multi-image input |
|---|---|---|---|
| OpenFlamingo v2 9B | Open-source | 9B parameters | Supported |
| Idefics2 8B | Open-source | 8B parameters | Supported |
| VILA 1.5 13B | Open-source | 13B parameters | Supported |
| LLaVA-NeXT-Interleave 7B | Open-source | 7B parameters | Supported |
| LLaVA-OneVision | Open-source | 7B parameters | Supported |
| Mantis 8B (clip-llama3) | Open-source | 8B parameters | Supported |
| Mantis 8B (siglip-llama3) | Open-source | 8B parameters | Supported |
| mPLUG-Owl3 7B | Open-source | 7B parameters | Supported |
| Deepseek-VL 7B chat | Open-source | 7B parameters | Supported |
| Pixtral 12B | Open-source | 12B parameters | Supported |
| GPT-4 Turbo | Proprietary | Not disclosed | Supported |
| GPT-4o | Proprietary | Not disclosed | Supported |
| Gemini Pro | Proprietary | Not disclosed | Supported |
| Claude 3.5 Sonnet | Proprietary | Not disclosed | Supported |

Table 4: Large vision language models evaluated on MRAG-BENCH.

- **Ground-Truth Text RAG**: Ground-truth textual descriptions are supplied where available, representing ideal textual knowledge.
- **Image plus Text RAG**: Some analyses provide both retrieved images and text to examine multimodal fusion.

Human participants are evaluated under analogous conditions, which allows direct comparison between human and model use of visually augmented knowledge.

# 5 Modeling Approach

## 5.1 Large Vision Language Models

MRAG-BENCH evaluates 14 large vision language models that accept multiple images as input. Four models are proprietary and ten are open-source. Table 4 summarizes their types and approximate scales.

All models are evaluated in their publicly released forms without fine-tuning on MRAG-BENCH. This design reflects the realistic use of general-purpose multimodal models.

## 5.2 Retrievers and Knowledge Sources

For the main experiments, CLIP is used as the multimodal image retriever. Given a query $Q$, CLIP embeds the query image and the text into a joint representation space and retrieves the most similar images from the corpus, yielding the list $I$ in Equation (1). CLIP is also used for text retrieval from the Wikipedia corpus by matching embeddings of queries and text passages.

In additional analyses, three more multimodal retrievers are evaluated: MagicLens, E5-V, and VISTA. Each retriever uses different architectures for multimodal representation learning. Their performances are compared using Recall@5, and their retrieved images are passed to a fixed large vision language model to study the relation between retriever quality and downstream accuracy.

### 5.3 Implementation Details

The implementation follows several standard practices.

- For each model, default decoding hyperparameters such as temperature, maximum answer length, and sampling strategy are taken from the official releases.
- Images are preprocessed according to each model's requirements, for example resolution normalization and color channel ordering.
- In retrieval-augmented settings, the top five retrieved images or text passages are used, except for the INCOMPLETE scenario where one exemplar is sufficient.
- Answers are extracted from model outputs using predefined parsing rules, with auxiliary prompts only in rare ambiguous cases.

The study does not propose a new model architecture. The contribution lies in the benchmark design and the systematic evaluation protocol.

## 6 Empirical Results

### 6.1 Overall Difficulty and Baseline Performance

MRAG-BENCH is a challenging benchmark. Without retrieval-augmented knowledge, the strongest model, GPT-4o, achieves an overall accuracy of 68.68%. With ground-truth image retrieval, GPT-4o reaches 74.50%, corresponding to an improvement of 5.82 percentage points.

Open-source models show lower performance. Without retrieval, their accuracies range from 26.83% to 53.29%. With ground-truth image retrieval, their accuracies range from 28.90% to 59.28%. These ranges indicate that ground-truth images provide useful information, but current open-source models still lag behind proprietary systems.

Human participants achieve 38.47% accuracy without retrieval. This relatively low baseline illustrates that many questions require specific visual knowledge beyond what is immediately visible in the query image. When humans are provided with retrieved images, accuracy increases to 61.38%. With ground-truth images, humans achieve 71.63%. Thus, humans gain 22.91 percentage points from retrieved images and 33.16 percentage points from ground-truth images.

The contrast between human and model gains is substantial. GPT-4o gains only 5.82 percentage points from ground-truth images, whereas humans gain 33.16 percentage points. This gap indicates that current large vision language models do not yet exploit retrieved images as effectively as human observers.

### 6.2 Effect of Retrieved and Ground-Truth Image Knowledge

All evaluated models benefit from ground-truth images, but the magnitude of improvement varies. Among open-source models, improvements from ground-truth images range from 2.07 to 11.31 percentage points. Among proprietary models, improvements range from 5.64 to 11.23 percentage points.

When noisy retrieved images from the CLIP retriever are used instead of curated ground-truth images, almost all open-source models experience accuracy drops. Their predictions are sensitive to misleading or partially relevant images. In contrast, proprietary models continue to gain accuracy with retrieved images, which suggests a stronger ability to identify and downweight noisy visual evidence.

A qualitative analysis shows that, in a fine-grained car recognition example, CLIP retrieves both correct exemplars and visually similar but incorrect images. Gemini Pro focuses on the correct exemplars and maintains the correct answer, while LLaVA-NeXT-Interleave is distracted by incorrect images and changes its answer from correct to incorrect. This example explains why open-source models are more frequently harmed by noisy retrieval.

## 6.3 Visual Knowledge versus Textual Knowledge

The study directly compares the benefits of visual and textual knowledge by evaluating LLaVA-NeXT-Interleave and GPT-4 Turbo under both image and text retrieval.

For LLaVA-NeXT-Interleave, retrieved images yield 40.35% accuracy, while retrieved text yields 37.99% accuracy. Ground-truth images yield 52.99% accuracy, and ground-truth text yields 41.09% accuracy. Therefore, visual knowledge outperforms textual knowledge by 2.36 percentage points under retrieval and by 11.90 percentage points under ground-truth conditions.

For GPT-4 Turbo, retrieved images yield 58.95% accuracy, and retrieved text yields 56.61%. Ground-truth images yield 62.85% accuracy, and ground-truth text yields 58.98%. Visual knowledge thus outperforms textual knowledge by 2.34 percentage points under retrieval and by 3.87 percentage points under ground-truth conditions.

When both ground-truth image and ground-truth text knowledge are provided, GPT-4 Turbo achieves further improvements beyond image-only conditions, whereas LLaVA-NeXT-Interleave does not consistently benefit from combining modalities. This observation suggests that stronger models can integrate multimodal evidence more effectively.

Overall, these results demonstrate that MRAG-BENCH is genuinely vision-centric. On this benchmark, retrieved images are more beneficial than retrieved text for both open-source and proprietary models.

## 6.4 Scenario-wise Performance Analysis

Scenario-wise analysis reveals that GPT-4o is the strongest single model across most scenarios. GPT-4o surpasses human performance on all perspective scenarios (ANGLE, PARTIAL, SCOPE, OCCLUSION) and on TEMPORAL and DEFORMATION scenarios among transformative cases. However, GPT-4o remains behind human performance in the INCOMPLETE and BIOLOGICAL scenarios.

The INCOMPLETE scenario is particularly challenging for models. Many models perform worse when additional images are provided, which implies that they do not systematically leverage visual exemplars for layout comparison and missing-part reasoning. Humans, by contrast, achieve 83.33% accuracy in the INCOMPLETE scenario when ground-truth images are available, indicating that humans can easily use complete exemplars to detect missing components.

## 6.5 Impact of Retriever Quality

The impact of retriever quality is studied by evaluating four multimodal retrievers: CLIP, MagicLens, E5-V, and VISTA. Each retriever is scored using Recall@5 on MRAG-BENCH, and its retrieved images are passed to LLaVA-NeXT-Interleave as the downstream model.

There is a strong positive correlation of approximately 95% between retriever Recall@5 and downstream accuracy. When a retriever more frequently places ground-truth images in its top five results, the downstream model attains higher accuracy. This confirms that retrieval-augmented reasoning strongly depends on retriever quality.

Despite similar Recall@5 scores for CLIP and VISTA, downstream accuracy still differs by about 2.07 percentage points. This suggests that not only the presence of relevant images but also their rank order and the composition of the retrieved set affect model predictions. This phenomenon is related to position bias in retrieval-augmented generation, which has been studied in text-only settings but remains largely unexplored in visual retrieval.

### 6.6 Number of Ground-Truth Image Examples

The default configuration uses five images per question for retrieval-augmented settings. MRAG-BENCH also investigates how performance changes with the number of ground-truth images.

For perspective and others scenarios, where an average of 20.4 ground-truth images is available per question, LLaVA-NeXT-Interleave is evaluated with 1, 2, 3, 5, 10, and 20 ground-truth images. The largest improvement occurs when moving from zero to one ground-truth image, which yields a gain of 5.64 percentage points. Performance continues to increase as more images are added and peaks at 10 images, which slightly outperforms using all 20 images.

This pattern suggests that there is an optimal range for the number of images. Too many images create a long visual context that is difficult for current models to process effectively. Furthermore, some questions only require one or a few high-quality exemplars, and additional images do not always provide incremental benefits.

## 7 Summary

### 7.1 Contributions and Novelty

The MRAG-BENCH study introduces a vision-centric benchmark and provides several methodological contributions.

- **Vision-centric benchmark design**: MRAG-BENCH focuses on scenarios where visual knowledge is more helpful than textual knowledge. In contrast to previous benchmarks, images are the primary retrieved modality.
- **Fine-grained scenario taxonomy**: The benchmark defines nine scenarios that cover perspective changes and transformative appearance changes, enabling fine-grained analysis of visual reasoning capabilities.
- **Systematic evaluation of models and humans**: Ten open-source and four proprietary large vision language models are evaluated together with human participants under multiple retrieval configurations, providing detailed comparisons between human and model use of visual knowledge.
- **Analysis of retrieval quality and context length**: The study quantifies the relationship between retriever Recall@5 and downstream accuracy and examines how performance varies with the number of ground-truth image examples.

### 7.2 Limitations and Future Directions

MRAG-BENCH also has limitations that point to future research directions.

- The benchmark focuses on static images and does not include video or three-dimensional scenes. Extending vision-centric retrieval evaluation to temporal and 3D data would better reflect many real-world tasks.
- All tasks use multiple-choice questions, which simplifies evaluation but does not capture open-ended generation or multi-step reasoning. Future work could incorporate free-form answers with structured evaluation metrics.

- The dataset draws heavily from specific image sources such as ImageNet, Flowers102, StanfordCars, GeoDE, and web images retrieved by keyword search. Some real-world object categories and environments remain underrepresented.
- Retrieval and fusion strategies are fixed in most experiments. Adaptive methods that select informative exemplars dynamically and arrange them in optimal order may further improve performance.
- Open-source models are more vulnerable to noisy retrieval than proprietary models. Developing training objectives and architectures that enhance robustness to misleading visual evidence is an important direction.

### 7.3 Concluding Remarks

MRAG-BENCH demonstrates that visually augmented knowledge is powerful but underutilized by current large vision language models. Humans achieve much larger performance gains from additional images than models do. Bridging this gap will require advances in retriever quality, visual context modeling, and multimodal reasoning algorithms. MRAG-BENCH provides a principled testbed and clear experimental targets for future research on retrieval-augmented multimodal systems.