

MAVIS: A Benchmark for Multimodal Source Attribution in Long-form Visual Question Answering

Seokwon Song¹, Minsu Park¹, Gunhee Kim^{1*}

¹Seoul National University
{seokwon.song, minsu.park}@vision.snu.ac.kr, gunhee@snu.ac.kr

Abstract

Source attribution aims to enhance the reliability of AI-generated answers by including references for each statement, helping users validate the provided answers. However, existing work has primarily focused on *text-only scenario* and largely overlooked the role of *multimodality*. We introduce MAVIS, the first benchmark designed to evaluate multimodal source attribution systems that understand user intent behind visual questions, retrieve multimodal evidence, and generate long-form answers with citations. Our dataset comprises 157K visual QA instances, where each answer is annotated with fact-level citations referring to multimodal documents. We develop fine-grained automatic metrics along three dimensions of informativeness, groundedness, and fluency, and demonstrate their strong correlation with human judgments. Our key findings are threefold: (1) LVLMs with multimodal RAG generate more informative and fluent answers than unimodal RAG, but they exhibit weaker groundedness for image documents than for text documents, a gap amplified in multimodal settings. (2) Given the same multimodal documents, there is a trade-off between informativeness and groundedness across different prompting methods. (3) Our proposed method highlights mitigating contextual bias in interpreting image documents as a crucial direction for future research.

Code — <https://github.com/seokwon99/MAVIS>

1 Introduction

Since the advent of open-domain generative models, detecting and mitigating potentially inaccurate or fabricated information has become a critical challenge (Ye et al. 2023; Zhang et al. 2023). This so-called hallucination problem arises from the models’ ability to generate highly fluent, human-like responses, making inaccuracies difficult to detect (Huang et al. 2025). This issue is particularly problematic in long-form generation, where the generated content frequently includes numerous pieces of information that are a mixture of true and false (Min et al. 2023).

These challenges suggest a growing need that not only generates factually accurate answers but also supports them with verifiable evidence. One such approach includes source attribution, in which systems enhance the verifiability of

*Corresponding author.

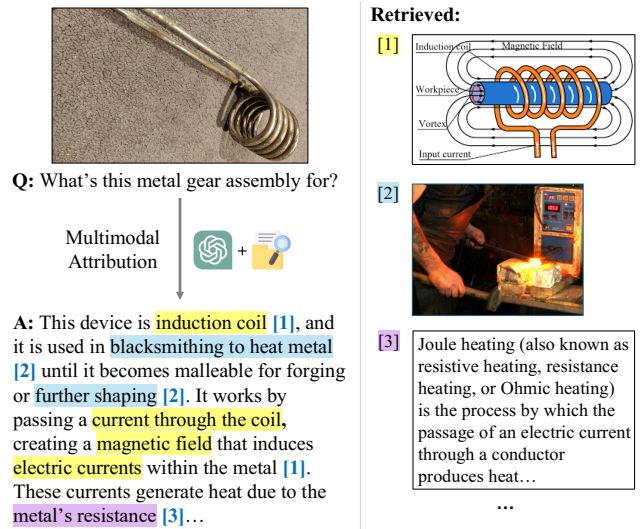


Figure 1: Example of the system’s response in our MAVIS benchmark. Given a user question paired with an input image, the system must generate a long-form answer supported by sentence-level citations that reference multimodal documents. Highlighted texts can be verified using the corresponding colored documents.

their long-form answers by including citations for each statement (Bohnet et al. 2022; Gao et al. 2023). Existing works in this area focus on the *text-only* setting, where systems interpret user queries expressed in natural language and retrieve evidence solely from textual sources.

However, text-only attribution may be insufficient, as many real-world scenarios require multimodal understanding. First, images are inherently *compact yet rich*, making them suitable for conveying dense and detailed information intuitively. For instance, as shown in document [1] of Figure 1, the image simultaneously represents multiple layers of information—such as “the structure of an induction coil”, “the generation of a magnetic field by current flowing through the coil”, and “the magnetic field inducing electric currents in metal”—in a compact and intuitive manner. Second, user-provided images play a critical role in understanding user intent and enabling precise responses. It is of

Dataset	# of Instance			Document		Answer Length	Document Modality	Fact-level Citation	Task Formulation
	Q	I	A	D_T/Q	D_I/Q				
ELI5 (2019)	272,000	0	272,000	1.0	0	130.6	Text	✗	Long-form QA
AquaMuse (2020)	5,519	0	5,519	6.0	0	105.9	Text	✗	Summarization
HowSumm (2021)	95,469	0	95,469	10.1	0	150.2	Text	✗	Summarization
WikihowQA (2023)	11,746	0	11,746	6.3	0	149.3	Text	✗	Long-form QA
LFRQA (2024)	26,907	0	26,907	3.0	0	76.3	Text	✓	Long-form QA & Text Retrieval
LONGFACT (2024b)	2,280	0	0	0	0	-	-	✗	Long-form QA
VizWiz-LF (2024)	600	600	4,200	0	0	41.2	-	✗	Long-form VQA
M2RAG (2024)	750	0	0	0 [†]	0 [†]	-	Text, Image	✗	Multimodal Generation
MRAMB-Bench (2025)	4,800	7,340	4,800	1.2	1.8	134.8	Text, Image	✗	Multimodal Generation
MAVIS (Ours)	81,173 (991)	67,140 (931)	157,586 (1,000)	5.9 (3.3)	3.1 (2.0)	76.5	Text, Image	✓	Long-form VQA & Multimodal Attribution

Table 1: Comparison of long-form question answering (LFQA) benchmarks. Q , I , and A denote the number of unique questions, images, and answers, respectively. D_T/Q and D_I/Q are the average number of text and image documents per question. The number in parentheses in our dataset indicates the statistics of the human-annotated set. The answer length is the average word count of answers. Fact-level citation indicates whether supporting documents are available for each verifiable fact. The dash (–) indicates that the corresponding feature is not covered in that benchmark. † denotes the absence of an annotated document for each question; however, an external knowledge base is utilized for retrieval.

ten difficult to grasp the intent of a question based solely on a natural language query—for example, “what’s this metal gear assembly for?”—without considering the accompanying image.

To bridge the gap between existing attribution benchmarks and real-world scenarios, we introduce MAVIS (Multimodal Attribution for Visual Question Answering), a benchmark designed to evaluate models on their ability to: (1) comprehend questions involving visual inputs, (2) generate effective search queries to retrieve multimodal documents, and (3) provide long-form answers with appropriate citations. We automatically construct a dataset of 157K instances, each containing a visual question and a ground-truth answer citing multimodal documents at the sentence level, and manually annotate 1K instances for evaluation. As shown in Table 1, our dataset uniquely includes both visual questions and multimodal supporting documents, and provides fact-level citation.

To ensure fine-grained evaluation of long-form responses and citation quality, we employ three metrics. (1) *Informativeness* measures how thoroughly the answer covers necessary information without unnecessary content; (2) *Groundedness* assesses how well the answer is supported by the citations; and (3) *Fluency* evaluates how fluent and coherent the output is. Our human evaluation results show that this automatic evaluation highly correlates with human judgments, making it a reliable evaluation method.

Through extensive experiments, we also find multiple intriguing observations.

1. Multimodal RAG generates more informative and fluent answers than unimodal RAG (i.e., text-only or image-only). However, LVLMs primarily rely on text when generating answers, making them less attentive to image-

based documents.

2. Given identical multimodal documents, there is a trade-off between informativeness and groundedness across different prompting methods, indicating the challenge of improving both simultaneously.
3. LVLMs fabricate information more frequently from image documents than from text documents. We introduce a knowledge extraction step before the final answer generation, effectively addressing this issue by mitigating contextual bias.

2 Related Works

Source attribution. Open-domain generative systems often produce plausible yet inaccurate content, known as hallucinations (Ye et al. 2023; Zhang et al. 2023). To mitigate this, attribution techniques have been introduced, enabling models to provide supporting evidence in the form of citations to improve verifiability (Gao et al. 2023; Li et al. 2024b; Sun et al. 2023; Slobodkin et al. 2024; Huang et al. 2024; Li and Ng 2024). However, these methods primarily address text-based inputs, with limited attention to the multimodal settings. We bridge this gap by adapting attribution approaches to multimodal contexts and identifying key factors for enhancing performance.

Long-form question answering (LFQA). LFQA datasets are widely used in AQA tasks, as long-form responses are prone to including inaccurate information (Liu, Zhang, and Liang 2023). ELI5 (Fan et al. 2019) and HowSumm (Boni et al. 2021) are popular LFQA datasets based on Reddit and WikiHow, respectively, and are evaluated by reference-based metrics such as ROUGE (Lin 2004) and BLEU (Papineni et al. 2002). To better address the open-ended nature of

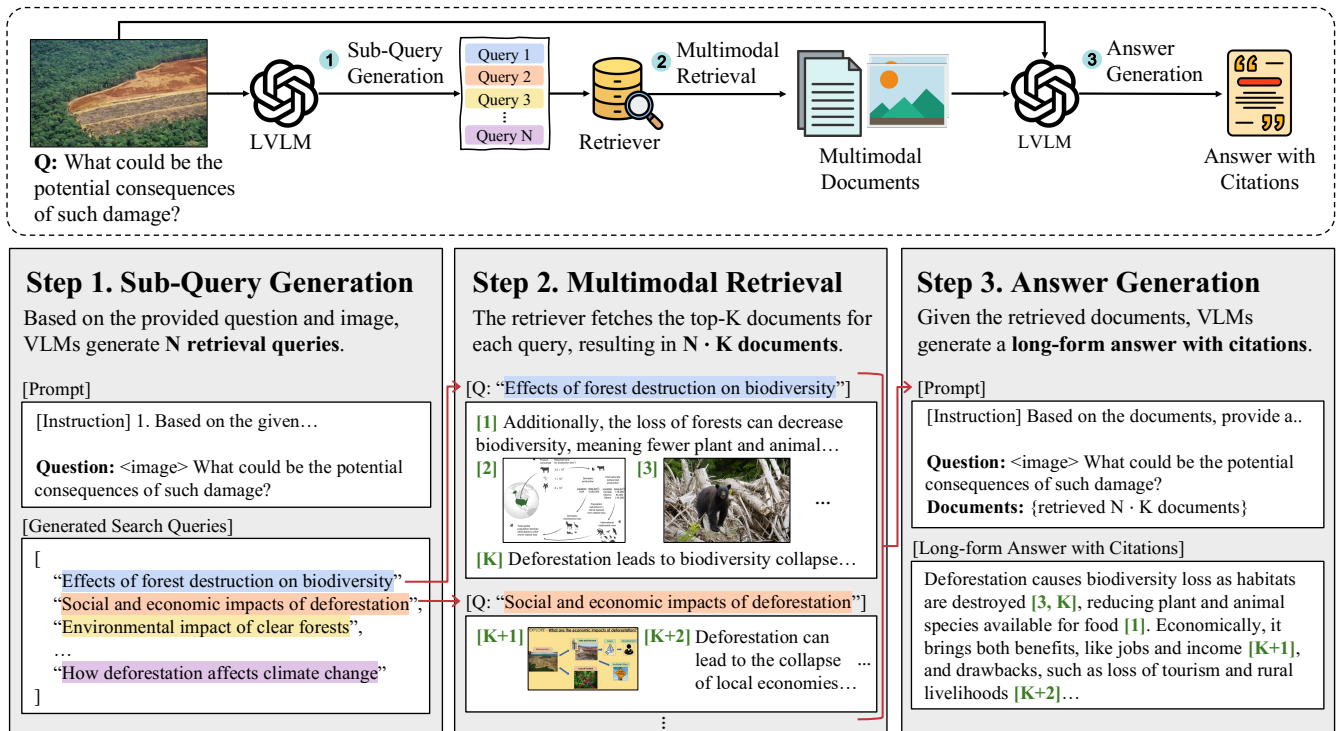


Figure 2: An illustration of the task formulation in our benchmark.

LFQA, LongFact (Wei et al. 2024b) verifies atomic claims via web search, while RAG-QA Arena (Han et al. 2024) employs pairwise preference evaluations. However, most existing datasets are still limited to text-only inputs, overlooking the role of multimodality. VizWiz-LF (Huh et al. 2024) introduces a long-form VQA task aimed at describing image content to blind or low-vision users. However, it does not consider grounding answers in external knowledge. M2RAG (Ma et al. 2024) and MRAMB-Bench (Yu et al. 2025) introduce multimodal retrieval-augmented generation to produce multimodal answers. However, the images included in these answers serve as part of the response rather than as evidence supporting the statements. Our work focuses on open-domain long-form VQA, which requires grounding answers in retrieved multimodal documents.

3 The MAVIS Benchmark

3.1 Task formulation

In our task, given a visual question consisting of a user question Q and an input image \mathcal{I} , a vision-language model (VLM) \mathcal{M} generate a long-form answer supported by verifiable evidence. Specifically, as illustrated in Figure 2, the task involves the following three steps:

- Search query generation:** Given Q and \mathcal{I} , \mathcal{M} generates a list of N search queries $\mathcal{S} = \{s_1, \dots, s_N\}$.
- Multimodal document retrieval:** The retriever \mathcal{R} fetches the top- K documents $\mathcal{D}_i = \{d_{i,j}\}_{j=1}^K$ for each search query s_i . In total, $N \cdot K$ documents are retrieved, denoted as $\mathcal{D} = \{D_1, \dots, D_N\}$.

- Answer generation with citations:** Given the retrieved \mathcal{D} , \mathcal{M} generates a final long-form answer with citations, each of which is enclosed in square brackets (e.g., [1]).

3.2 Dataset collection process

As discussed in §3.1, our benchmark involves two essential components: a visual question and multimodal supporting documents. We automatically (1) collected long-form VQA data from user forums, and (2) retrieved supporting documents for each answer in fact-level. Subsequently, we (3) manually annotated a test set for reliable evaluation. More details about the dataset collection can be found in Appendix A.

Step 1. Collection of long-form VQA data

Raw data. We collect 6M posts containing both images and comments from Reddit pushshift dumps from 2005-06 to 2023-12. We filter them by three rules: the title must be a question starting with a question word, ending with a question mark, and comments must be long (over 500 characters, 3+ sentences). By treating titles as questions and comments as answers, we obtain 432,817 VQA instances.

Visual dependency. We filter out instances that can be answered without an input image, following Chen et al. (2024a). We instruct LLMs to answer the question without the input image and evaluate their answers against the forum answers. We utilize four LLMs—GPT-4o (Hurst et al. 2024), LLaMA-3.3-70B (Dubey et al. 2024), Mixtral-8x7B (Jiang et al. 2024), and Phi-4-14B (Abdin

Type	Model	F1-score	
		Text	Image
Textual	NLIDeBERTaV3-184M (2021)	41.4	-
	FlanT5Verifier-11B (2024)	40.7	-
	Qwen3-8B (2025)	53.7	-
Visual	OFA-VE-470M (2022)	-	23.1
	SkyworkVLReward-8B (2025)	-	50.1
Multi	Qwen2VL-7B (2024)	45.0	49.6

Table 2: Model performances in verifying the groundedness of each sentence on the retrieved documents across different modalities. The F-1 scores are measured by a small set of ground truths that the authors manually label whether each document supports the corresponding fact. Due to computational constraints, we use about 11B open-source models.

et al. 2024)—and employ InternVL-2.5-38B (Chen et al. 2024b) as the evaluator. VQA instances correctly answered by at least one LLM are removed, resulting in 157,586 VQA instances. The prompts used for both LLMs and evaluator models are provided in Appendix A.1.

Step 2. Collection of supporting multimodal documents

Atomic fact extraction. A long-form answer typically contains multiple pieces of information (Min et al. 2023; Jing et al. 2023). We extract atomic facts from the ground-truth answer using GPT-4.1 with more details in Appendix A.2. On average, each answer contains 4.3 atomic facts.

Collection of multimodal documents. To collect external multimodal documents grounding atomic facts, we considered two search sources: Google Programmable Search, which searches the entire web for relevant images, and the Colossal Clean Crawled Corpus (C4) (Raffel et al. 2020), a collection of hundreds of gigabytes of English text scraped from the web. Using each atomic fact as a query, we retrieved the top five passages from each source.

Automatic document filtering. Each VQA instance initially contains an average of 43 multimodal documents (4.3 facts \times 2 modalities \times 5 documents), which may overwhelm annotators during the filtering process. Hence, we automatically remove irrelevant documents using entailment models (EMs). Table 2 shows that Qwen3-8B (Yang et al. 2025) and SkyworkVLReward-8B (Wang et al. 2025) perform the best for text and image entailment, respectively. Using these models, we retain 930K text documents and 500K image documents. Further details are provided in Appendix A.3.

Step 3. Human annotation

Sampling. We sample a subset from the automatically collected dataset for human annotation. First, to ensure multimodal knowledge is necessary for answering the visual question, we select instances containing more than two atomic facts supported by multimodal documents. Second, to mitigate domain bias in the test set, we perform balanced sampling across different domains, resulting in 1,000

VQA instances. Detailed information is provided in Appendix A.4.

Criteria. Annotators label the sampled instances based on two criteria: (1) Are the atomic facts relevant to the question? (2) Are the facts accurately supported by the documents? As a result, we find that 85.7% of atomic facts are relevant to the question. Among the relevant atomic facts, 87.3% are supported by their corresponding documents. Finally, we annotate 1K test instances, comprising 3K relevant atomic facts supported by 5.1K ground-truth multimodal documents. Notably, all of our experimental results are based on this human-annotated test set. Further details on the annotation including inter-annotator agreement are in Appendix A.5.

3.3 Human verification process

To demonstrate the effectiveness of our data construction process, we involve human verifiers to assess whether the questions in our dataset (1) genuinely seek information or advice, as opposed to merely sharing information, advertising, or making statements, and (2) whether an attached image is necessary to understand or answer the user’s question, or unnecessary or irrelevant. We compare 50 visual questions from our dataset with their corresponding raw Reddit posts, each binary-labeled by two MTurk workers, with average inter-annotator agreement (IAA, measured by Cohen’s Kappa) of 0.74 and 0.68, respectively.

As a result, 89% of the questions in our dataset exhibit information-seeking intent, compared to 49% of the raw posts. Regarding image dependency, 89% of questions in our dataset depend on attached images, whereas 68% of raw posts require one. This confirms that our data construction process effectively removes irrelevant instances from the raw sources.

4 Evaluation Metrics

We evaluate long-form answers based on three aspects: (1) informativeness, (2) groundedness, and (3) fluency. We use GPT-4.1 as the evaluator for groundedness and informativeness, and MAUVE (Pillutla et al. 2021) for fluency. Further details are provided in Appendix B.

4.1 Informativeness

Long-form answers can be paragraph-length responses that should be helpful and comprehensive. However, they often contain a large amount of information, making binary judgments challenging (Min et al. 2023). To address this, we define two sub-metrics for *informativeness*. (1) **Completeness** measures how thoroughly the model’s answer \mathcal{A} covers the necessary information from the GT answer \mathcal{G} , and (2) **Relevance** assess whether \mathcal{A} contains only information relevant to the user question \mathcal{Q} and the user-provided image \mathcal{I} .

1. **Completeness.** For the essential information that the answer \mathcal{A} should cover, we use atomic facts $\{g_1, \dots, g_m\}$, extracted from \mathcal{G} and filtered by human annotators. Given a fact g_j and the model’s answer \mathcal{A} , the completeness score $c(g_j, \mathcal{A})$ measures how thoroughly g_j is addressed

by \mathcal{A} as $\{1: \textit{fully relevant}, 0.5: \textit{partially relevant}, 0: \textit{not relevant}\}$. The final completeness score is the average of $c(g_j, \mathcal{A})$ across all g_j :

$$\text{Completeness}(\mathcal{A}) = \frac{1}{m} \sum_{j=1}^m c(g_j, \mathcal{A}).$$

- Relevance.** Models should not generate excessive or irrelevant information to achieve high completeness. Given the model’s answer $\mathcal{A} = \{a_1, \dots, a_n\}$, user question \mathcal{Q} , and input image \mathcal{I} , the relevance score $r(a_i, \mathcal{Q}, \mathcal{I})$ indicates how appropriately each answer sentence a_i addresses \mathcal{Q} and \mathcal{I} as $\{1: \textit{fully relevant}, 0.5: \textit{partially relevant}, 0: \textit{not relevant}\}$. The final relevance score is computed as the average relevance across all a_i :

$$\text{Relevance}(\mathcal{A}) = \frac{1}{n} \sum_{i=1}^n r(a_i, \mathcal{Q}, \mathcal{I}).$$

4.2 Groundedness

We evaluate citation quality in terms of answer groundedness, using two sub-metrics. (1) **Recall** measures whether the answer is fully supported by citations, and (2) **Precision** identifies redundant or irrelevant citations. To do this, we first pair each sentence a_i in the model’s answer \mathcal{A} with its corresponding cited documents \mathcal{C}_i based on citation numbers. Each a_i is then assigned a supportedness score $s(\cdot, \cdot)$ based on how well a_i is supported by the documents as $\{1: \textit{fully relevant}, 0.5: \textit{partially relevant}, 0: \textit{not relevant}\}$.

- Recall.** We assess how well each answer sentence $a_i \in \mathcal{A}$ is supported by its cited documents \mathcal{C}_i . The recall score for \mathcal{A} is the average of $s(a_i, \mathcal{C}_i)$ across all a_i :

$$\text{Recall}(\mathcal{A}) = \frac{1}{n} \sum_{i=1}^n s(a_i, \mathcal{C}_i).$$

- Precision.** We evaluate how relevant every cited document is with its answer. For each answer sentence a_i , let $\mathcal{C}_i = \{c_{i,1}, \dots, c_{i,m}\}$ denote the set of its cited documents. The precision score for the answer \mathcal{A} is the average, across all sentences a_i , of the mean supportedness scores between a_i and each citation in \mathcal{C}_i :

$$\text{Precision}(\mathcal{A}) = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{|\mathcal{C}_i|} \sum_{c_{i,j} \in \mathcal{C}_i} s(a_i, c_{i,j}) \right).$$

4.3 Fluency

To measure how fluent and human-like the model’s answer \mathcal{A} is, we adopt MAUVE (Pillutla et al. 2021), as done in (Gao et al. 2023). Fluency mainly serves as a sanity check, ensuring MAUVE scores remain sufficiently high.

5 Experiments

5.1 Models

Large vision language models. We select four state-of-the-art LVLMs. For proprietary models, we use (1) Claude-3.5-Sonnet-20241022, and (2) GPT-4o-240806 (Hurst et al.

Retriever	NDCG@10	Recall@100
Fine-tuned on WebQA (Chang et al. 2022)		
CLIP-DPR	0.1567	0.4355
UniVL-DR	0.1136	0.3244
MARVEL-DPR	0.1292	0.4188
MARVEL-ANCE	0.1322	0.3948
Fine-tuned on ClueWeb (Overwijk et al. 2022)		
MARVEL-DPR	0.1098	0.4357
MARVEL-ANCE	0.1460	0.4398
Fine-tuned on M-BEIR (Wei et al. 2024a)		
MM-Embed		
+ text-seeking query	0.2216	0.5909
+ image-seeking query	0.2217	0.6074
+ averaged query embedding	0.2565	0.6977

Table 3: Multimodal retrieval performance on the human annotated test set.

2024). For public models, we use (3) LLaVa-OneVision-Qwen2-72b-ov-hf (Li et al. 2024a), and (4) QwenVL-72B-Instruct (Wang et al. 2024). Implementation details are explained in Appendix C.1.

Multimodal retrievers. We build a large-scale multimodal database of 2.5M documents—1.4M from our data and 1.1M (389K images, 787K texts) from the WebQA corpus (Chang et al. 2022). We evaluate several multimodal retrievers, including CLIP-DPR (Liu et al. 2022), UniVL-DR (Liu et al. 2022), MARVEL (Zhou et al. 2023), CLIP-SF (Wei et al. 2024a), and MM-Embed (Lin et al. 2024). For each test instance, we generate four search queries using the aforementioned LVLMs and measure retrieval accuracy against the GT supporting documents. As shown in Table 3, the modality-aware retriever MM-Embed shows the best performance using averaged query embeddings for both text and image queries. Thus, we adopt MM-Embed as our default retriever.

5.2 Baselines

Retrieval modalities. We evaluate three RAG settings based on the modality of the knowledge base: Text-RAG, Image-RAG, and Multi-RAG. In Text-RAG and Image-RAG, the retriever selects documents from the corresponding unimodal database. In Multi-RAG, the retriever selects documents from a combined text-image database.

Answer generation. We explore three answer generation methods. Instructions are detailed in Appendix C.2.

- Vanilla prompting:** We prompt each model to generate answers with inline citations. This end-to-end approach enables the simultaneous generation of sentences and their corresponding citations.
- Chain-of-Thought (CoT) prompting:** Previous studies (Slobodkin et al. 2024; Berchansky et al. 2024) adopt Chain-of-Thought (CoT) prompting (Wei et al. 2022) to enhance the accuracy of attributions. We utilize a guided

	Evaluation Metrics (%)							Statistics			
	Informativeness			Groundedness			Fluency	Retrieved		Utilized	
	F1-score	Complete	Relevant	F1-score	Recall	Precision	MAUVE	Text	Image	Text	Image
LLaVa-OneVision											
+ Text-RAG	31.9	26.0	78.0	66.7	70.3	65.5	80.4	5.0	0.0	3.2	0.0
+ Image-RAG	36.2	28.2	87.0	17.0	19.9	20.3	85.6	0.0	5.0	0.0	2.8
+ Multi-RAG	33.3	27.6	81.5	62.7	66.1	61.6	88.7	3.3	1.7	2.2	0.3
Qwen2.5VL											
+ Text-RAG	30.7	23.4	77.8	77.9	79.8	77.0	60.1	5.0	0.0	3.4	0.0
+ Image-RAG	30.7	24.2	75.1	49.2	50.0	54.0	81.5	0.0	5.0	0.0	3.0
+ Multi-RAG	32.3	25.4	77.3	72.0	72.5	73.9	62.8	2.8	2.2	2.0	1.1
Claude-3.5-Sonnet											
+ Text-RAG	29.6	24.0	68.6	79.1	80.9	78.1	69.7	5.0	0.0	3.5	0.0
+ Image-RAG	32.3	28.5	68.9	53.1	52.2	57.7	71.5	0.0	5.0	0.0	3.2
+ Multi-RAG	35.4	29.3	75.6	74.7	75.2	76.0	72.9	2.5	2.5	2.1	1.1
GPT-4o											
+ Text-RAG	37.0	30.4	82.7	73.4	76.9	71.3	69.7	5.0	0.0	3.1	0.0
+ Image-RAG	40.2	38.0	87.4	37.1	38.4	45.4	71.5	0.0	5.0	0.0	2.7
+ Multi-RAG	44.4	42.5	85.4	62.1	65.0	63.2	73.3	2.9	2.1	2.1	0.6

Table 4: Performance of LVLMS over modalities of knowledge base. Each baseline uses Vanilla prompting under the single retrieval setting ($N = 1, K = 5$). **Bold numbers** indicate the best performance. We calculate the F1-score for groundedness and informativeness as representative values.

Method	Inf.	Grd.	Flu.
	F1-score	F1-score	MAUVE
LLaVa-OneVision			
+ Vanilla	33.3	62.7	88.7
+ Vanilla + KE (Ours)	35.2	63.0	79.8
Qwen2.5VL			
+ Vanilla	32.3	72.0	50.8
+ Vanilla + KE (Ours)	34.9	72.0	65.9
+ CoT	38.6	68.2	72.4
+ CoT + KE (Ours)	39.0	70.4	69.1
Claude-3.5-Sonnet			
+ Vanilla	35.4	74.7	72.9
+ Vanilla + KE (Ours)	37.5	75.1	96.2
+ CoT	39.6	64.6	63.2
+ CoT + KE (Ours)	40.4	65.0	94.5
GPT-4o			
+ Vanilla	44.4	62.1	73.3
+ Vanilla + KE (Ours)	44.8	66.4	70.1
+ CoT	32.7	69.3	79.9
+ CoT + KE (Ours)	36.5	70.0	81.2

Table 5: Performance of LVLMS for each answer generation method under single-query retrieval setting ($N = 1, K = 5$). All baselines use the Multi-RAG setting. **Bold numbers** indicate the best performance.

reasoning framework consisting of: (1) finding relevant documents from the given set, (2) extracting relevant information from each document, and (3) generating the final answer using the relevant information.

- **Knowledge Extraction (KE) step:** LVLMS may refer-

ence documents but fabricate interpretations to generate wrong plausible answers. To mitigate this, we introduce an additional step prior to answer generation, independently extracting knowledge from documents without using the user question Q or input image I , thereby preventing potential bias introduced by these inputs. Given documents \mathcal{D} , we extract factual information $\tilde{\mathcal{D}} = \{\mathcal{M}(\mathcal{I}_{\text{extract}}, d) \mid d \in \mathcal{D}\}$ using a knowledge extraction instruction $\mathcal{I}_{\text{extract}}$, and then the VLM \mathcal{M} generates a final long-form answer with citations from $\tilde{\mathcal{D}}$, while groundedness is evaluated with the original documents.

5.3 Results for single-query retrieval

We report automatic evaluation results in the single-query retrieval setting ($N = 1$), where the Top-5 documents ($K = 5$) are retrieved. Table 4 compares performance across different retrieval modalities. Table 5 explores different answer generation methods in the Multi-RAG setting.

Utilizing multimodal documents improves informativeness and fluency. In Table 4, we compare retrieval modalities to identify the most effective knowledge modality for MAVIS. All models, except LLaVa-OneVision, demonstrate increased informativeness when using multimodal retrieval. For instance, Multi-RAG with GPT-4o achieves a 7.4% higher F1-score in informativeness compared to Text-RAG, and a 4.2% higher score than Image-RAG. Additionally, while all baselines show good fluency overall, Multi-RAG shows comparatively better fluency than uni-RAG, except for Qwen2.5VL.

LVLMS struggle to ground their answers in image documents. In Table 4, Image-RAG performs the lowest groundedness, with F1-scores ranging from 17.0% to 53.1%,

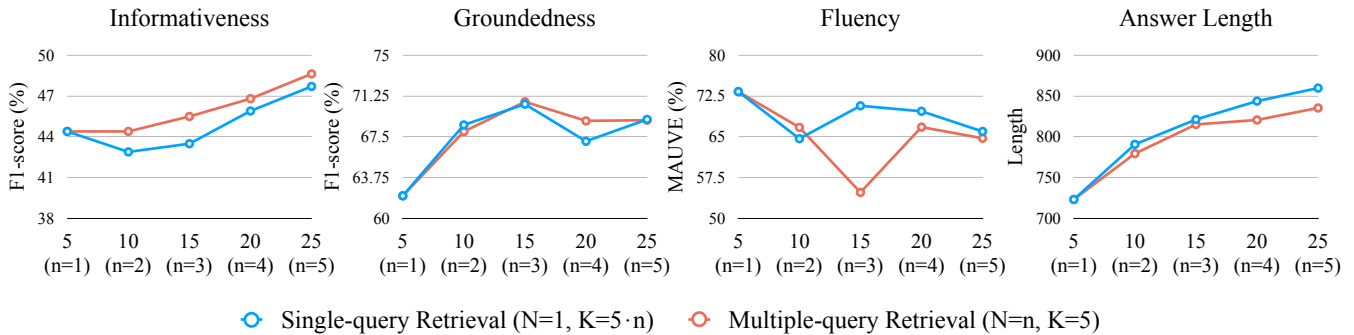


Figure 3: Performance of Multi-RAG with GPT-4o using single-query ($N = 1$) and multiple-query ($N = n$) retrieval. All baselines use Vanilla prompting. The x -axis indicates the total number of retrieved documents.

whereas Text-RAG attains the highest groundedness with F1-scores from 66.7% to 73.4%. This indicates that LVLMS struggle to generate accurate citations when referencing image documents compared to text documents. Multi-RAG shows a 4.0%–11.3% decrease in groundedness relative to Text-RAG; this reduction is further analyzed in §5.6.

Trade-off between informativeness and groundedness across prompting methods. In Table 5, we compare two prompting methods—Vanilla and Chain-of-Thought (CoT)—in Multi-RAG, and observe a trade-off between informativeness and groundedness. For instance, CoT prompting with Qwen2.5VL increases informativeness by 6.3% but decreases groundedness by 3.8% compared to Vanilla prompting. Conversely, GPT-4o with CoT prompting increases groundedness by 7.2% while decreasing informativeness by 11.7% compared to Vanilla prompting. These results highlight the inherent difficulty of simultaneously improving informativeness and groundedness.

Knowledge extraction can mitigate this trade-off. As shown in Table 5, knowledge extraction improves one metric (informativeness or groundedness) without sacrificing performance on the other. For GPT-4o with Vanilla prompting, after applying the knowledge extraction (KE) step, groundedness and informativeness increase by 4.3% and 0.4%, respectively. Similarly, GPT-4o with CoT prompting exhibits improvements of 4.2% in informativeness and 0.7% in groundedness after applying KE. Further analysis of the knowledge extraction step is provided in §5.6.

5.4 Results of retrieving more documents

In Figure 3, we present automatic evaluation results for retrieving more documents (≥ 5) using Vanilla prompting in Multi-RAG. We compare two baseline strategies while retrieving the same number of ($n \cdot k$) documents. (1) Single-query retrieval: a single query ($N = 1$) retrieves $K = n \cdot k$ documents, and (2) multiple-query retrieval: $N = n$ queries are generated, each retrieving $K = k$ documents.

Effects of retrieving more documents. Both informativeness and answer lengths generally increase with the number of retrieved documents. Specifically, retrieving 25 documents yields up to a 4.2% increase in informativeness com-

Metric	Sub-metric	Pearson Corr.
Groundedness	Recall	0.903
	Precision	0.803
Informativeness	Completeness	0.733
	Relevance	0.855

Table 6: Pearson correlation between human and GPT-4.1 scores for each metric.

pared to retrieving 5 only. Groundedness also improves with more documents, peaking at 15 documents with a 8.4% gain, beyond which additional documents offer no further benefit and slightly reduce performance. However, fluency shows a downward trend as more documents are retrieved, suggesting that integrating multiple documents may negatively impact the naturalness of model responses.

Effects of multiple-query retrieval. The multiple-query retrieval can incorporate diverse information. Thus, it achieves greater informativeness compared to single-query retrieval. However, multiple-query retrieval exhibits less stable fluency, indicating that increased document diversity may negatively affect response fluency.

5.5 Results for human evaluation

We conduct a human study to validate automatic evaluation results with GPT-4.1. We collect 100 answers each from Qwen2.5-VL-72B and GPT-4o across three RAG frameworks (Image-RAG, Text-RAG, Multi-RAG) using Vanilla prompting, totaling 600 answers. Annotators rate answers on recall $s(a_i, C_i)$, precision $s(a_i, C_i)$, completeness $c(g_j, \mathcal{A})$, and relevance $r(a_i, Q, \mathcal{I})$. As shown in Table 6, Pearson correlation coefficients exceed 0.733 for all criteria, confirming GPT-4.1’s reliability in the automatic evaluation of long-form answers. The annotation guidelines, inter-annotator agreement, and detailed results are provided in Appendix B.2.

5.6 Results for modality-specific evaluation

To investigate the capability of LVLMS to utilize documents for each modality, we conduct a modality-specific evaluation

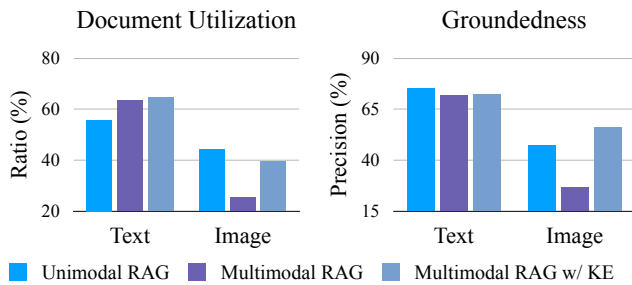


Figure 4: Comparison of per-modality groundedness and document utilization between Unimodal-RAG (Text-RAG and Image-RAG), and Multimodal-RAG with and without knowledge extraction (KE). Each metric is averaged over the four LVLMs described in §5.1.

using two document-centric metrics: groundedness and the document utilization ratio. Groundedness is measured using precision rather than recall to assess the model’s ability to utilize each document. The document utilization ratio per modality is defined as the average ratio of the number of used documents to that of retrieved documents.

How do LVLMs handle multimodal knowledge? In Figure 4, we analyze the changes of each modality’s contribution from a unimodal to a multimodal setting by comparing Multi-RAG with Uni-RAG (Text-RAG and Image-RAG). In Multi-RAG, text groundedness slightly decreases from 75.2% to 71.9%, whereas image groundedness substantially drops from 46.9% to 26.8%. Additionally, text utilization increases from 55.9% to 63.5%, while image utilization notably decreases from 44.0% to 25.5%. These results indicate that compared to Uni-RAG, Multi-RAG tends to primarily rely more on textual documents when generating answers, resulting in less attention to image documents compared to Uni-RAG.

Which modality would benefit from knowledge extraction? We propose a knowledge extraction step to mitigate the issue of LVLMs that fabricate to generate plausible answers. To identify which document modality benefits the most, we compare Multi-RAG and Multi-RAG w/ KE baselines. Significant improvements are observed for image documents, with groundedness rising from 26.8% to 56.0%, and document utilization from 25.5% to 39.3%. This indicates that LVLMs frequently fabricate information from image documents, and knowledge extraction effectively addresses this issue by mitigating contextual biases during interpretation.

To what extent do text documents affect attention to image documents? We found that LVLMs’ reliance on text reduces their attention to image documents. To investigate this further, we conducted a controlled experiment, shown in Figure 5, to examine how the number of text documents affects attention to images. First, when a single text document is added—compared to when no text documents are present—image utilization drops significantly, from 58.2% to 26.25%. This suggests that even one text document, which

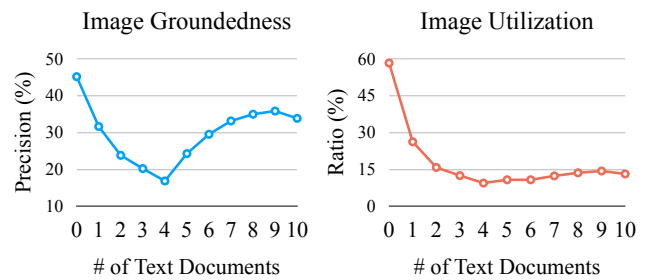


Figure 5: Effect of adding text documents on image attention in GPT-4o. We fix the number of image documents at 5 in a single-query setting ($N = 1$), and the x-axis represents the number of additional retrieved text documents. Documents are given in random order.

is small relative to the number of images, can cause GPT-4o to generate more text-focused responses. Second, when the number of text documents exceeds four, image utilization remains roughly constant at around 15%, but image groundedness increases substantially, from 16.9% to 33.85%. These results indicate that, as recent studies (Deng et al. 2025; Wu et al. 2025) have reported, LVLMs exhibit text dominance and underutilize image documents; however, increasing the number of text documents beyond a certain point improves the accuracy of image document grounding.

6 Future Directions

In our experiments, we focus on prompting LVLMs without updating their model weights. By releasing sentence-level multimodal attribution data, we leave the exploration of optimizing source attribution in multimodal scenarios. While our work focuses on text and image documents, models could also leverage other modalities, such as video or audio, to ground their answers. For instance, video is particularly suitable for understanding dynamic events, motion, or sequences. Extending the dataset to include diverse modalities is an important direction for future work.

7 Conclusion

We introduce MAVIS, a benchmark for evaluating visual question answering with multimodal attribution. It includes 157K automatically generated instances and 1K manually annotated for evaluation. Experiments show that multimodal RAG yields more fluent, informative answers than unimodal RAG, but also reveals key limitations. LVLMs often over-rely on text, leading to weaker visual grounding. We also found a trade-off between informativeness and groundedness: higher grounding reduces informativeness, and vice versa. Our analysis highlights that stronger contextual bias with image documents than text, suggesting that an explicit knowledge extraction step can help mitigate this issue.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP)

grant funded by the Korea government (MSIT) (No. RS-2019-II191082, SW StarLab, No. RS-2022-II220156, Fundamental research on continual meta-learning for quality enhancement of casual videos and their 3D metaverse transformation, No. RS-2025-02263841, Development of a Real-time Multimodal Framework for Comprehensive Deepfake Detection Incorporating Common Sense Error Analysis, and No. RS-2021-II211343, Artificial Intelligence Graduate School Program, Seoul National University), and Sovereign AI Foundation Model Project (Data Track), organized by the Ministry of Science and ICT (MSIT) and supported by the National Information Society Agency (NIA) (2025-AI Data-wi43).

References

- Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Berchansky, M.; Fleischer, D.; Wasserblat, M.; and Izsak, P. 2024. CoTAR: Chain-of-Thought Attribution Reasoning with Multi-level Granularity. *arXiv preprint arXiv:2404.10513*.
- Bohnet, B.; Tran, V. Q.; Verga, P.; Aharoni, R.; Andor, D.; Soares, L. B.; Ciaramita, M.; Eisenstein, J.; Ganchev, K.; Herzig, J.; et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.
- Bolotova-Baranova, V.; Blinov, V.; Filippova, S.; Scholer, F.; and Sanderson, M. 2023. WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5291–5314.
- Boni, O.; Feigenblat, G.; Lev, G.; Shmueli-Scheuer, M.; Sznajder, B.; and Konopnicki, D. 2021. Howsumm: A multi-document summarization dataset derived from wikipow articles. *arXiv preprint arXiv:2110.03179*.
- Chang, Y.; Narang, M.; Suzuki, H.; Cao, G.; Gao, J.; and Bisk, Y. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16495–16504.
- Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.
- Deng, A.; Cao, T.; Chen, Z.; and Hooi, B. 2025. Words or Vision: Do Vision-Language Models Have Blind Faith in Text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3867–3876.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints, arXiv:2407*.
- Fan, A.; Jernite, Y.; Perez, E.; Grangier, D.; Weston, J.; and Auli, M. 2019. ELI5: Long form question answering. *arXiv preprint arXiv:1907.09190*.
- Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.
- Han, R.; Zhang, Y.; Qi, P.; Xu, Y.; Wang, J.; Liu, L.; Wang, W. Y.; Min, B.; and Castelli, V. 2024. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. *arXiv preprint arXiv:2407.13998*.
- He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.
- Huang, C.; Wu, Z.; Hu, Y.; and Wang, W. 2024. Training language models to generate text with citations via fine-grained rewards. *arXiv preprint arXiv:2402.04315*.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.
- Huh, M.; Xu, F.; Peng, Y.-H.; Chen, C.; Murugu, H.; Gurari, D.; Choi, E.; and Pavel, A. 2024. Long-Form Answers to Visual Questions from Blind and Low Vision People. *arXiv preprint arXiv:2408.06303*.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Jing, L.; Li, R.; Chen, Y.; and Du, X. 2023. FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models. *arXiv preprint arXiv:2311.01477*.
- Kulkarni, S.; Chammas, S.; Zhu, W.; Sha, F.; and Ie, E. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Li, D.; Sun, Z.; Hu, B.; Liu, Z.; Hu, X.; Liu, X.; and Zhang, M. 2024b. Improving attributed text generation of large language models via preference learning. *arXiv preprint arXiv:2403.18381*.
- Li, J.; and Ng, H. T. 2024. Think&Cite: Improving Attributed Text Generation with Self-Guided Tree Search and Progress Reward Modeling. *arXiv preprint arXiv:2412.14860*.

- Li, M.; Peng, B.; Galley, M.; Gao, J.; and Zhang, Z. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- Lin, S.-C.; Lee, C.; Shoeybi, M.; Lin, J.; Catanzaro, B.; and Ping, W. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.
- Liu, N. F.; Zhang, T.; and Liang, P. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.
- Liu, Z.; Xiong, C.; Lv, Y.; Liu, Z.; and Yu, G. 2022. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *arXiv preprint arXiv:2209.00179*.
- Ma, Z.-A.; Lan, T.; Tu, R.-C.; Hu, Y.; Zhu, Y.-S.; Zhang, T.; Huang, H.; Wu, Z.; and Mao, X.-L. 2024. Multi-modal Retrieval Augmented Multi-modal Generation: Datasets, Evaluation Metrics and Strong Baselines. *arXiv preprint arXiv:2411.16365*.
- Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.
- Overwijk, A.; Xiong, C.; Liu, X.; VandenBerg, C.; and Callan, J. 2022. Clueweb22: 10 billion web documents with visual and semantic information. *arXiv preprint arXiv:2211.15848*.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; and Harchaoui, Z. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34: 4816–4828.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.
- Sanyal, S.; Xiao, T.; Liu, J.; Wang, W.; and Ren, X. 2024. Are Machines Better at Complex Reasoning? Unveiling Human-Machine Inference Gaps in Entailment Verification. *arXiv preprint arXiv:2402.03686*.
- Slobodkin, A.; Hirsch, E.; Cattan, A.; Schuster, T.; and Dagan, I. 2024. Attribute first, then generate: Locally-attributable grounded text generation. *arXiv preprint arXiv:2403.17104*.
- Sun, H.; Cai, H.; Wang, B.; Hou, Y.; Wei, X.; Wang, S.; Zhang, Y.; and Yin, D. 2023. Towards verifiable text generation with evolving memory and self-reflection. *arXiv preprint arXiv:2312.09075*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, 23318–23340. PMLR.
- Wang, X.; Wang, P.; Pei, J.; Shen, W.; Peng, Y.; Hao, Y.; Qiu, W.; Jian, A.; Xie, T.; Song, X.; et al. 2025. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*.
- Wei, C.; Chen, Y.; Chen, H.; Hu, H.; Zhang, G.; Fu, J.; Ritter, A.; and Chen, W. 2024a. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 387–404. Springer.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Wei, J.; Yang, C.; Song, X.; Lu, Y.; Hu, N.; Huang, J.; Tran, D.; Peng, D.; Liu, R.; Huang, D.; et al. 2024b. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.
- Wu, H.; Tang, M.; Zheng, X.; and Jiang, H. 2025. When Language Overrides: Revealing Text Dominance in Multimodal Large Language Models. *arXiv preprint arXiv:2508.10552*.
- Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Ye, H.; Liu, T.; Zhang, A.; Hua, W.; and Jia, W. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.
- Yu, Q.; Xiao, Z.; Li, B.; Wang, Z.; Chen, C.; and Zhang, W. 2025. MRAMG-Bench: A Comprehensive Benchmark for Advancing Multimodal Retrieval-Augmented Multimodal Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3616–3626.
- Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.
- Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023. Siren’s song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.
- Zhou, T.; Mei, S.; Li, X.; Liu, Z.; Xiong, C.; Liu, Z.; Gu, Y.; and Yu, G. 2023. MARVEL: unlocking the multi-modal capability of dense retrieval via visual module plugin. *arXiv preprint arXiv:2310.14037*.