FM²DS: Few-Shot Multimodal Multihop Data Synthesis with Knowledge Distillation for Question Answering

Anonymous ACL submission

Abstract

Multimodal multihop question answering (MMQA) requires reasoning over images and text from multiple sources, an essential task for many real-world applications. Despite advances in visual question answering, this multihop setting remains underexplored due to a lack of quality datasets. Existing methods focus on single-hop, single-modality, or short texts, limiting real-world applications like interpreting educational documents with long, multimodal content. To fill this gap, we introduce *FM²DS*, the first framework for creating a high-quality dataset for MMQA. Our approach consists of a 5-stage pipeline that involves acquiring relevant multimodal documents from Wikipedia, synthetically generating high-level questions and answers, and validating them through rigorous criteria to ensure data quality. We evaluate our methodology by training models on our synthesized dataset and testing on two benchmarks: MultimodalQA and WebQA. Our results demonstrate that, with an equal sample size, models trained on our synthesized data outperform those trained on human-collected data by 1.9 in exact match (EM) score on average. Additionally, we introduce *M²QA-Bench* with 1k samples, the first benchmark for MMQA on long documents, generated using *FM²DS* and refined by human annotators.

1 Introduction

Multimodal multihop question answering (MMQA) involves answering complex questions by integrating information from text, images, and tables. In real-world applications such as interpreting medical documents, this challenge is amplified by the need to reason over long, multimodal content. Current methodologies in MMQA typically leverage in-context learning methods, prompting LVLMs to retrieve relevant information from multimodal sources (Tejaswi et al., 2024) and then perform reasoning (Yang et al., 2023). However, these models often demand significant computational resources due to their large parameter counts, making them costly to deploy even during inference (Ye et al., 2024). This limitation emphasizes the need for more efficient frameworks that can operate effectively with minimal annotated data. A practical solution is to use a smaller model capable of both retrieving the necessary information from sources and performing reasoning. This can be achieved by fine-tuning the model on a MMQA dataset. Existing datasets often rely on short snippets or repetitive templates, limiting generalizability to complex settings with long texts and multiple modalities (Chang et al., 2021; Talmor et al., 2021; Jiang et al., 2024; Chen et al., 2024a). Additionally, creating new similar datasets is challenging, requiring extensive human annotation (Lu et al., 2022a; Chen et al., 2023a).

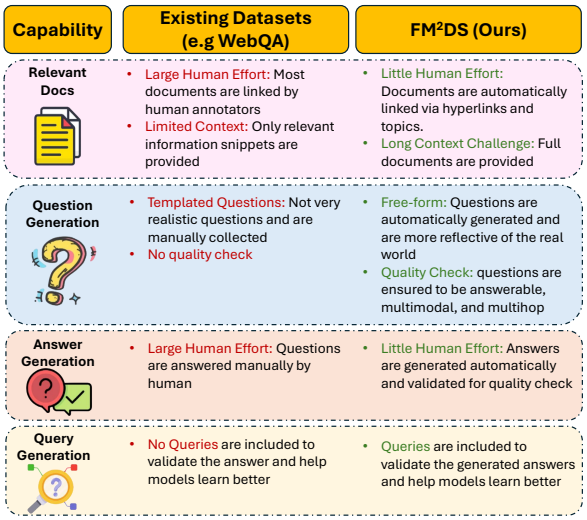

Figure 1: Unlike traditional datasets that rely on human annotators, templates, or snippets, *FM²DS* is fully automated, using long documents as sources and applying validation to ensure questions are answerable, multimodal, and multihop.

In this work, we propose *FM²DS*, a novel

data synthesis framework designed specifically for MMQA over **long documents**. Our approach synthesizes MMQA data from documents that are interconnected through various relationships, such as thematic similarities or sequential events. This framework leverages naturally occurring document relationships and requires minimal hand-crafted data, thereby broadening the range of reasoning types used in question generation.

As illustrated in Figure 1, *FM*^{2}*DS* enables the generation of non-templated question-answer pairs based on full documents rather than brief information snippets. The data generated by our method *FM*^{2}*DS* includes **query** component - **a step-by-step guide for retrieving relevant information from multiple documents** - enabling smaller models trained on this synthesized data to learn how to tackle complex questions in a manner similar to larger models. This methodology allows users to create a custom MMQA dataset with fewer than ten human-annotated samples, thereby facilitating the fine-tuning of smaller LVLMs for specific applications.

FM^2*DS* leverages Wikipedia's extensive knowledge base and hyperlink structure to select document pairs with shared topical relevance or hyperlink connections and prompt LVLMs to perform question generation, question answering, and query generation. We incorporate validation steps to enhance the quality of the generated data and discard any outputs that are factually incorrect. Through empirical evaluation on established MMQA benchmarks, we show that *FM*^{2}*DS* significantly improves model performance, achieving on average a 1.9 exact match (EM) score improvement across two benchmarks: MultimodalQA and WebQA.

Our key **contributions** are: **(I)** introducing a new framework for synthesizing high-quality MMQA training data for VLMs; **(II)** using a robust validation pipeline to ensure data quality; **(III)** introducing a challenging MMQA benchmark requiring reasoning over multiple modalities and sources; and **(IV)** showing that models fine-tuned on our synthetic data outperform those trained on human-labeled datasets, advancing MMQA while reducing manual effort.

2 Related Work

Within the Question Answering (QA) literature, synthesis of training data has been predominantly focused on unimodal (text-only) scenarios. We review various similar works that have established the foundation for our work in few-shot data synthesis.

Unimodal Data Synthesis Synthetic data is increasingly used for model training. He et al. (2022) show that combining labeled and synthetic text from language models (LMs) improves NLP performance. Entire synthetic datasets have also been created for tasks like classification (Tsui, 2024), with Li et al. (2023) demonstrating GPT-3.5's effectiveness in generating reliable classification data. Similarly, Chen et al. (2024b) show that synthetic data can significantly boost small models on multi-hop QA with minimal human annotation.

Multimodal Data Synthesis Research on multimodal data synthesis with LVLMs remains limited, with most efforts focused on generating new data from model's pre-trained knowledge. Zhang et al. (2024) synthesize abstract images with reasoning tasks, while Mehta et al. (2024) generate multimodal data using unimodal models for pre-training. In MMQA, Wu et al. (2024) propose SMMQG, which uses multimodal RAG to generate questions from short snippets, focusing on multimodality rather than multihop reasoning. In contrast, *FM*^{2}*DS* uses full multimodal documents, resulting in a more challenging dataset with diverse multihop questions that better reflect real-world tasks. Moreover, while SMMQG is confined to predefined question types, *FM*^{2}*DS* enables large-scale generation and supports knowledge distillation for smaller models through step-by-step queries that guide complex multi-document reasoning.

3 Proposed Method: *FM*^{2}*DS*

Our five-stage pipeline for *FM*^{2}*DS* (Figure 2) synthesizes high-quality multimodal QA pairs. It begins by grouping documents via topic matching and Wikipedia hyperlinks, followed by few-shot sample selection, question synthesis, answer generation, and query construction, each with their built-in validation. See Appendix I for examples.

3.1 Stage 1: Creating a Pool of Related Documents

We collect relevant documents from Wikipedia using the WikiWeb2M dataset (Burns et al., 2023), which includes nearly 2 million pages. Documents are linked via two methods: hyperlinks and latent topics identified through multimodal topic

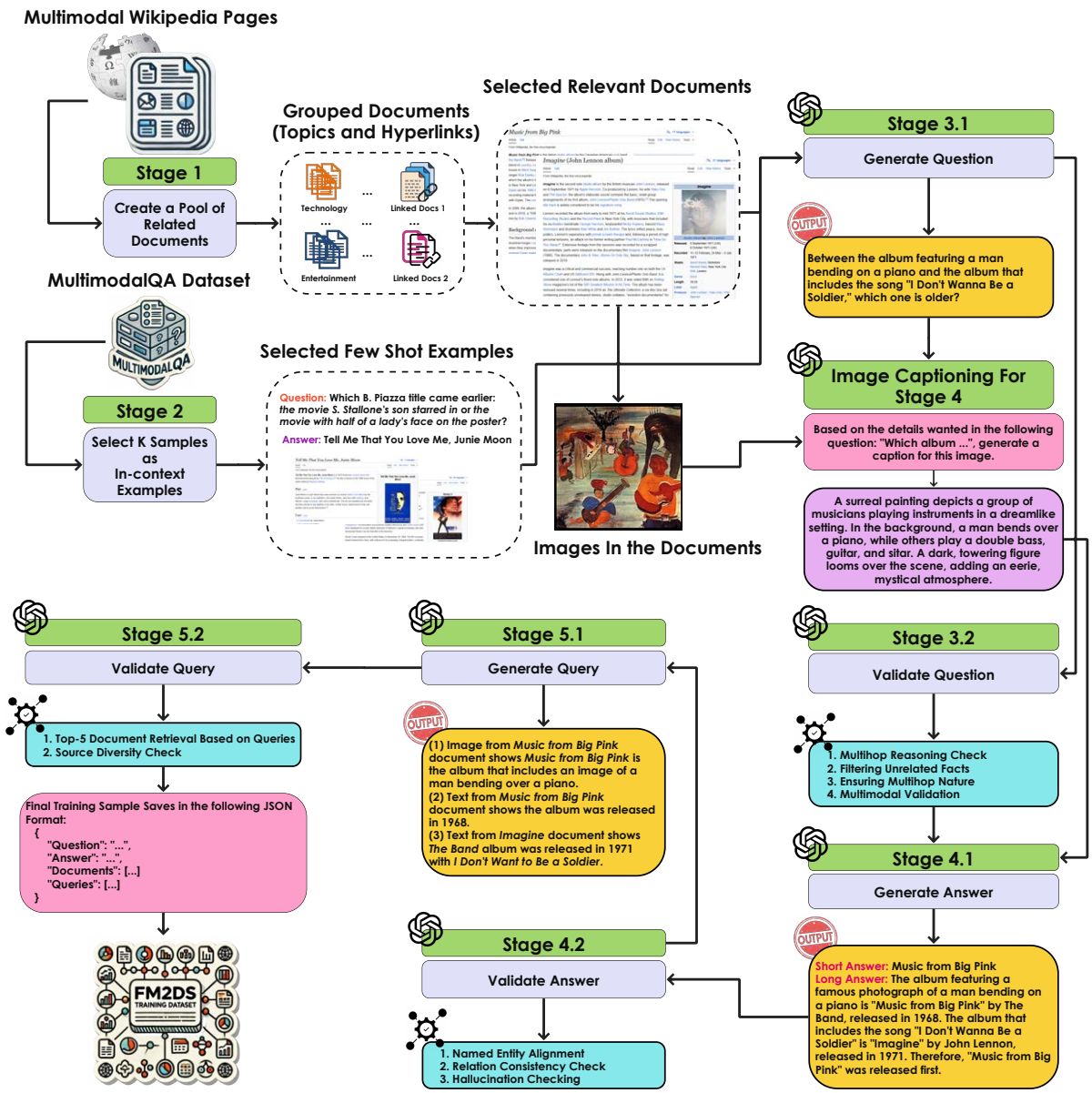

Figure 2: The FM2DS pipeline consists of five stages for generating high-quality multihop multimodal QA samples. In Stage 1, a pool of related Wikipedia documents is retrieved by leveraging topic similarity and hyperlink connections to ensure contextual richness. Stage 2 selects few-shot in-context examples from the MultiModalQA dataset (Talmor et al., 2021) to guide generation. Stage 3 focuses on question generation (3.1) and validation (3.2), ensuring questions require multihop reasoning, are answerable, and grounded in both text and images. Stage 4 generates (4.1) and validates (4.2) answers through named entity alignment, relation consistency, and hallucination checks. Finally, Stage 5 generates (5.1) and validates (5.2) retrieval queries to collect diverse and relevant supporting documents. The resulting samples are saved in a structured format for use in MMQA training and evaluation.

modeling with the Multimodal-Contrast model (González-Pizarro and Carenini, 2024). Since Multimodal-Contrast can not handle long documents, we split each document into shorter segments containing at most one image, apply topic modeling to each segment, then merge the results and remove duplicates. This combination captures both clear and subtle relationships across documents, integrating textual and visual information.

3.2 Stage 2: Creating Few-Shot Samples

We sample multihop questions from the MultiModalQA dataset (Talmor et al., 2021), which requires reasoning across text, images, and tables. As our samples are based on full documents rather than short information snippets like in MultimodalQA, we crawled the complete Wikipedia HTML pages using the entity links provided in MultimodalQA, which are associated with the dataset's examples.

We then compiled few-shot samples using these full HTML pages—complete with images and tables—paired with their corresponding questions. We randomly select up to three samples for question generation in our experiments.

3.3 Stage 3: Question Generation and Validation

Question Generation We use GPT-4-turbo (OpenAI et al., 2024) to generate multihop, multimodal questions from few-shot samples based on MultiModalQA few-shot examples. Due to context limitations, inputs are limited to grouped sets containing 2 or 3 documents. Our prompt (Appendix A) is designed to ensure that questions require reasoning across all documents and at least two modalities, avoiding unrelated combinations such as: "How did Einstein contribute to relativity and when was Princeton established?"—a question that spans multiple documents, but lacks meaningful multihop reasoning.

Question Validation Our framework includes validation stages to ensure questions meet multihop and multimodal criteria. While the model was prompted to avoid simple concatenations, we further evaluated this aspect.

We used LLama-3.1-8B (Abhimanyu Dubey et al., 2024) to decompose questions and check if parts could be answered with a single document. If all parts of the question were answerable with a single document, we discarded such question that include unrelated facts (see *Unrelated Facts* example in Table 1). Otherwise, we retained only the the parts of the questions that required information from multiple documents to ensure the revised question met the multihop criteria. However, a potential issue was that, even when the facts were related, the questions could still become open-ended, requiring explanations or combined answers (see example *Related Facts, Open-ended* in Table 1). In order to follow the standard of question answering, and make the evaluation process easier, we used GPT-4o to rephrase the question without conjunctions while maintaining its multihop nature, resulting in *Concise Multihop Question* (Table 1).

Another key step in validation was ensuring the questions were truly multimodal. After verifying that a question was multihop, we tested whether it remained answerable when the documents were limited to a single modality (e.g., text-only, image-only, or table-only). Using GPT-4o (refer to Section 3.4 for details), we checked if the question could be answered with just one modality. If so, we discarded it, as it failed to meet the multimodal requirement. This step helped refine the dataset to include only questions that genuinely required reasoning across multiple modalities and documents.

3.4 Stage 4: Answer Generation and Validation

Answer Generation We used GPT-4o to generate concise answers from multiple documents, including text and images. The model was instructed to provide a long answer and a short answer with only key information and no extra explanation. To help the model focus on specific details of images in the given documents to answer the multimodal question, we include question-related captions for the images. For example, if the question asks about the geometric shapes in an image (see Figure 15), the model generates a caption describing the shapes. This makes it easier for the model to answer the question accurately.

Answer Validation We validated answers using named entity recognition (NER) and relation extraction, following prior work (Rajpurkar et al., 2018; Fabbri et al., 2022). NER ensured key entities and numbers in the answer matched the documents, while relation extraction verified that entity relationships were consistent with the source content (via Spacy (Wu and He, 2019)). For including image content, we used the same question-related caption generated by GPT-4o (e.g., noting a building's color if relevant to the question) similar to answer generation. To reduce hallucinations, we prompted GPT-4o five times and accepted answers only if all outputs (5/5) agreed. To evaluate the effectiveness of our answer validation process, we conducted a human study to assess the quality of the filtered questions and answers. The results of this evaluation are presented in Section 6.

3.5 Stage 5: Query Generation and Validation

Query Generation We generate queries using GPT-4o based on the question-answer pairs and related documents to enhance retrieval effectiveness. These queries guide the smaller model trained on FM2DS-generated data to retrieve specific and relevant information, improving its ability to answer questions accurately. By narrowing down the content, we can extract key details such as named entities, relationships, and contextual cues aligned

Question	Type
In what year did Mike Tyson become the youngest heavyweight champion, and who is the president of the United States?	Unrelated Facts
Question	**Type**
In what year did Mike Tyson become the youngest heavyweight champion, and who was the president of the United States at that time?	Related Facts, Open-ended
Question	**Type**
Who was the president of the United States when Mike Tyson became the youngest heavyweight champion?	Concise Multihop Question

Table 1: Examples of factual questions with varying degrees of relevance and conciseness, demonstrating progression from unrelated to concise multihop reasoning.

with the question. This targeted approach ensures that the generated answers are not only concise and accurate, but also directly grounded in evidence from the documents.

Query Validation To validate the queries, we used MuRAG (Chen et al., 2022), which encodes text and images into a shared embedding space for multimodal retrieval. For each generated query, we retrieved the top-5 documents retrieved by MuRAG. If more than one of the original source documents used to generate the question appeared in the top-5, the query was considered well-formed. This process ensures the query effectively captures diverse, relevant information and can help teach smaller models how to retrieve supporting evidence for answering questions.

4 Proposed Benchmark: *M²QA-Bench*

We introduce *M²QA-Bench*, a benchmark of 1k diverse Q&A pairs to evaluate LVLMs on complex MMQA with full documents. Unlike templated datasets (Talmor et al., 2021), questions are varied and challenging (see Appendix G for details on complexity and diversity). Answering requires cross-modal reasoning and information extraction from full documents, including images and tables. See Table 2 for key statistics (more in Appendix G) and Appendix I for samples generated by *FM²DS* for *M²QA-Bench*. To create this benchmark, we used the *FM²DS* pipeline to generate 1,200 samples, which were verified by three annotators for correctness, multihop reasoning, multimodality, and answer accuracy. Each sample was scored 1 (valid) or 0 (invalid). This annotation required minimal human effort (2.2 min/question on average) due to structured queries. Samples averaging below 0.75 were removed, leaving 1,142 (i.e removing only 5% of the total); we then randomly selected 1,000 for the benchmark to ensure consistency in evaluation and reduce potential sampling bias. Annotator agreement (**Fleiss' Kappa** (Fleiss, 1971)) was **0.83**.

Statistic	Value
Image Modality Percentage	73.6%
Table Modality Percentage	89.6%
Both Image and Table Modality Percentage	63.6%
Average Question Length (Word)	23.77
Average Answer Length (Word)	1.95
Average Source Documents Per Question	2.29

Table 2: Key statistics of the proposed multimodal multihop question answering benchmark.

5 Experiments and Results

This section compares our synthesized dataset to human-annotated ones. All experiments used *one in-context example* during synthesis (see Appendix C for effects of varying number in-context examples). GPT-4o was used in the pipeline (Appendix D shows results with other LVLMs). Models were evaluated using Exact Match (EM) for accuracy and F1 for partial match quality. Further experimental details can be found in Appendix B.

5.1 Comparison with Human-Annotated Datasets

Unlike prior methods like MultiModalQA (Talmor et al., 2021) and WebQA (Chang et al., 2021), our approach is fully automated with no human involvement (minimal human evaluation was used in creating the *M²QA-Bench* only). This section compares the quality of our synthesized data against these human-annotated datasets. We trained LLaVA-1.6 (Liu et al., 2023b,a, 2024), InternVL-2 (Chen et al., 2023b, 2024c), and Idefics-2 (Laurençon et al., 2023, 2024b) on varying sizes of WebQA and MultiModalQA, evaluating on their respective test sets. We also trained the same models on *FM²DS*-generated training data and evaluated them on the same test sets to assess the effectiveness of the syn-

Model	Test Dataset									
	MultiModalQA					WebQA				
		EM		F1			EM		F1	
	FT (Real/Syn)	Real	Syn	Real	Syn	FT (Real/Syn)	Real	Syn	Real	Syn
LLaVa-1.6-7B	5k/5k	64.61	**69.68**	73.13	**76.52**	5k/5k	69.88	**75.49**	78.27	**87.61**
LLaVa-1.6-7B	10k/10k	73.96	**75.14**	78.36	**79.48**	10k/10k	77.49	**79.06**	82.59	**82.76**
LLaVa-1.6-7B	23.8k/21k	78.79	**79.41**	82.35	80.65	34.2k/16k	81.36	**82.48**	**85.79**	82.83
LLaVa-1.6-13B	10k/10k	77.45	**79.46**	79.12	**81.32**	10k/10k	80.22	**83.24**	84.26	**84.95**
LLaVa-1.6-13B	23.8k/21k	82.95	**83.56**	83.71	**84.76**	34.2k/13k	83.36	**85.34**	86.92	**87.64**
InternVL-2-8B	5k/5k	69.43	**73.92**	79.77	**84.27**	5k/5k	78.19	**81.27**	87.64	**90.21**
InternVL-2-8B	10k/10k	76.42	**77.25**	86.42	**87.06**	10k/10k	83.86	**85.34**	91.82	**94.05**
InternVL-2-8B	23.8k/17k	81.36	**82.95**	**90.2**	89.24	34.2k/15k	85.67	**86.58**	88.05	**92.23**
InternVL-2-26B	10k/10k	78.79	**79.23**	**88.91**	88.44	10k/10k	85.76	**86.49**	92.23	**93.19**
InternVL-2-26B	23.8k/16k	84.6	**85.29**	89.76	**91.24**	34.2k/15k	86.36	**87.74**	**91.82**	90.21
Idefics-2-8B	5k/5k	67.19	**71.48**	77.42	**79.32**	5k/5k	75.34	**78.31**	84.51	**86.92**
Idefics-2-8B	10k/10k	75.4	**76.57**	**85.39**	83.71	10k/10k	80.11	**83.86**	88.18	**90.07**
Idefics-2-8B	23.8k/18k	81.37	**81.85**	89.12	**89.76**	34.2k/15k	87.49	**87.67**	91.82	**92.23**
LLaVa-1.6-7B	None	50.85		56.34		None	56.37		65.44	
LLaVa-1.6-13B	None	56.83		61.17		None	61.79		68.32	
InternVL-2-8B	None	61.27		68.26		None	68.01		76.39	
InternVL-2-26B	None	**68.39**		**74.08**		None	**74.59**		**80.61**	
Idefics2-8B	None	60.47		66.41		None	64.79		72.38	
GPT-4o	None	**83.56**		**87.91**		None	**86.49**		**90.23**	
Llama-3.2-90B	None	77.18		80.37		None	82.22		86.9	
Claude-3.5-Sonnet	None	73.84		77.38		None	76.29		79.43	

Table 3: Comparison of synthetic and human-annotated data across various models. Smaller models were evaluated with 5k, 10k, and full training sets (23.8k for MultiModalQA, 34.2k for WebQA), while larger models used 10k and full sets. For each model, the listed synthetic sample size is the smallest (divisible by 1k) that outperforms the same model trained on the full human-labeled set. For real samples, we used the WebQA training set for testing on the WebQA test set, and similarly for MultiModalQA. "None" indicates default pretrained models. For more results, see Appendix E.

thesized data in comparison to human-annotated datasets.

As shown in Table 3, models trained on *FM*^{2}*DS* data outperform those trained on original datasets, despite using longer documents. WebQA seems easier than MultiModalQA, with better performance from fewer samples. On average, EM improved by 1.81 for MultiModalQA and 1.96 for WebQA using equal or fewer synthetic samples. While EM gains often lead to higher F1, some cases show F1 drops due to hallucinated answers reducing string overlap. Models trained on fewer synthetic samples from *FM*^{2}*DS* match the performance of those trained on full datasets, showing faster convergence (see Section 5.2). Larger models also perform better with the same synthetic data—e.g., LLaVA-1.6-13B vs. 7B. GPT-4o leads among large LVLMs, likely due to its role in data generation (Refer to Appendix J for qualitative analysis). When comparing these human annotated data with our synthesized data, a common question is: "Why not paraphrase existing datasets instead of synthesizing new ones?". The answer is paraphrasing does not enable domain adaptation, and as shown in Appendix K, fine-tuning on our synthetic data outperforms training on a paraphrased version of MultiModalQA.

5.2 Learning Efficiency Comparison

To evaluate learning efficiency, we ran experiments with InternVL-2-8B using incremental training sizes from 1k to 10k (in 1k steps) on both synthetic and human-annotated data. For real data, we used the same dataset for training and testing—e.g., when testing on WebQA, training samples were taken from the WebQA training set.

As shown in Figure 3(a & b), our synthesized data outperforms real data at smaller training sizes,

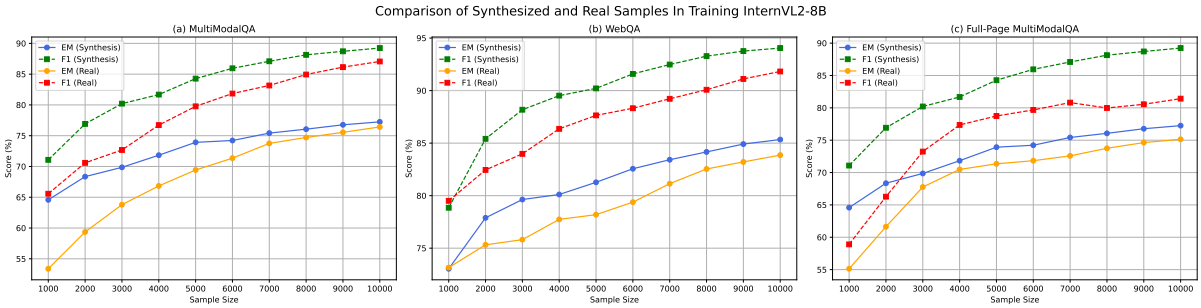

Figure 3: ***(a) and (b)***: EM and F1 comparison on 1k–10k samples for InternVL-2-8B shows that *FM*^{2}*DS*'s synthetic data outperforms human-annotated data, with the gap narrowing as sample size increases. ***(c)***: Similar comparison using full Wikipedia pages from the MultimodalQA dataset to match our synthetic data format.

Training Set	Test Set			
	MultiModalQA		Ours	
	EM	F1	EM	F1
MultiModalQA	63.2	74.24	48.6	61.45
Ours	**68.4**	**77.88**	**55.8**	**69.06**

Table 4: Cross-Dataset Evaluation Results With InternVL-2-8B for MultiModalQA and Our Synthesized Benchmark.

Steps Included	Test Dataset			
	MultiModalQA		WebQA	
	EM	F1	EM	F1
-	62.78	64.39	69.32	78.56
Question Validation	64.61	70.12	70.28	78.44
+ Answer Validation	67.96	73.56	74.59	79.67
+ Query	70.83	75.84	77.49	81.86
+ Query Validation	**73.92**	**84.27**	**81.27**	**90.21**

Table 5: Results of the *FM*^{2}*DS* with and without key steps like query generation and verification. All other steps are included in all of the results. The plus sign ("+") at the start each steps means all the previous steps were included as well.

though the gap narrows as sample size grows. Near 10k samples, learning efficiency with synthetic data declines more than with real data—likely due to its broader knowledge coverage. While this diversity aids early learning, it can lead to saturation, unlike real data, which offers more focused patterns and sustains steady learning (Hong et al., 2023; Maini et al., 2024).

In a related experiment on MultiModalQA, we used full Wikipedia pages via linked articles as training data instead of information snippets. This was not possible for WebQA, as its source links mostly point to WikiMedia pages with limited text. Figure 3(c) shows that models trained on full Wikipedia pages initially achieve better improvement per 1k samples; however, this advantage declines after approximately 3k samples. This suggests that, while full-page real data offers some early benefits over short snippets, it still lacks the generality and consistency of our synthesized dataset. In contrast, the synthesized data—with its built-in queries and diverse content—continues to support steady learning, likely due to its broader coverage and higher quality.

5.3 Cross-Dataset Evaluation

To assess the generalizability of our synthesized data, we conducted a cross-dataset evaluation using InternVL-2-8B. We used 1k samples from *M*^{2}*QA-Bench* and 1k from the MultiModalQA test set, both with full Wikipedia pages as sources. The model was trained separately on 5k samples from our dataset and 5k from MultiModalQA, using the same input format. As shown in Table 4, the model trained on our data outperformed the one trained on MultiModalQA across both test sets, demonstrating stronger generalization. This also reflects the greater complexity and diversity of our benchmark, compared to MultiModalQA's template-based questions. See Appendix H to see the performance on model trained on *FM*^{2}*DS*'s data when using on a out of domain MMQA dataset.

5.4 Key Stages in *FM*^{2}*DS*

To assess the impact of each data filtering and validation step, we evaluated InternVL-2-8B on 5k training samples under different configurations. As shown in Table 5, each step contributed to performance gains. Question validation improved relevance by filtering questions tailored to the complex, multimodal, multihop requirements of the test sets, and reduced hallucinations, boosting F1 with more accurate answers. Answer validation removed in-

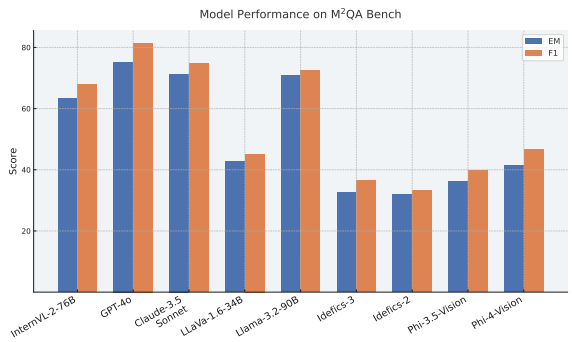

Figure 4: Bar plot showing EM and F1 scores of multimodal models on *M²QA-Bench*.

correct samples, while query generation distilled knowledge from larger models, improving EM and F1. Query validation reinforced consistency by ensuring proper structure. Refer to Appendix L for statistics like **"Rejection Rate"** on the validation stages used to remove specific samples during data synthesis.

5.5 *M²QA-Bench* Evaluation Results

To assess the complexity of *M²QA-Bench* and compare model performance, we evaluated a diverse set of models, as shown in Figure 4. GPT-4o stands out, outperforming even larger models like LLaMa-3.2-90B and Claude-3.5-Sonnet by a notable margin. Interestingly, smaller models in the 4B–8B range, particularly those from the Phi family (Abdin et al., 2024a,b), also achieve competitive results despite their scale. GPT-4o's advantage may stem in part from its involvement in the initial data generation, but this remains an open question, as human annotators thoroughly reviewed and corrected the dataset to ensure fairness, reduce bias, and maintain high quality.

6 Human Evaluation of Answer Validation

To evaluate the accuracy and impact of our automatic answer validation component, we conducted a human evaluation study. After generating questions for 100 samples, we continued the pipeline using two methods: *FM²DS* with and without answer validation. These samples were divided into four batches of 25, with each batch evaluated by three participants to mitigate potential bias or human errors. In total, 12 participants contributed through our evaluation platform (see Appendix F), assessing which method produced the correct answer.

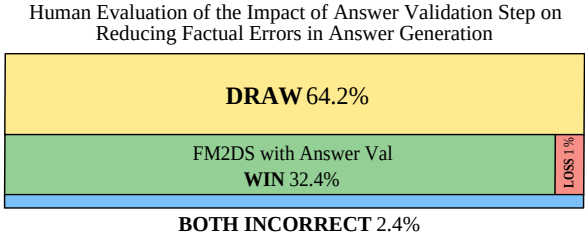

Figure 5: Human evaluation results, more people preferred *FM²DS* with answer validation (green region) than without (red region).

Participants were instructed to verify the correctness of each answer by reviewing the associated Wikipedia pages. Once they determined the accuracy of each answer, they were asked to select one of four options: (1) Method 1's answer is correct, (2) Method 2's answer is correct, (3) both methods generated the correct answer, or (4) neither answer is correct (Method 1 and Method 2 were randomly assigned to the datasets generated with and without answer validation, respectively.). The results, shown in Figure 5, reveal a clear trend: answer validation increases the likelihood of correct answers, even though the model without validation still produced correct answers in over 60% of cases. The **Fleiss' Kappa** was **0.91**—indicating strong agreement—though this is expected, as the task involved factual questions with definitive answers rather than subjective judgments.

7 Conclusion & Future Works

We present *FM²DS*, a novel methodology for synthesizing high-quality data for multimodal multihop question answering. Unlike existing approaches that are limited to single-hop and single-modality settings, *FM²DS*generates complex QA pairs that require reasoning over multiple modalities and sources, with minimal human intervention. Our framework enables the creation of a large-scale dataset that significantly boosts model performance, surpassing models trained on human-curated data in terms of test accuracy. These results demonstrate the effectiveness of synthetic data for advancing the state of the art in multimodal multihop QA. Additionally, *FM²DS*offers a scalable and efficient solution for training data-hungry language models. For future work, we plan to synthesize MMQA samples using sources beyond Wikipedia, including multilingual content, code snippets, videos, and other diverse information type.

Limitations

While *FM*^{2}*DS* offers a robust pipeline for synthesizing high-quality multimodal multihop QA data, several limitations remain.

First, although the framework incorporates strong validation steps—including factual consistency checks, named entity alignment, and hallucination detection—it is not immune to errors. Subtle factual inaccuracies and hallucinations may still persist, especially in answers grounded in complex visual content. Despite using multiple generations and automated agreement checks, there is still a risk that some incorrect samples pass through undetected.

Second, our reliance on large-scale generative models such as GPT-4o throughout multiple stages—including question generation, answer synthesis, captioning, and validation—makes the pipeline computationally expensive. This cost is further amplified by the need to regenerate failed samples that do not pass intermediate validation steps. In some settings, particularly when generating large-scale datasets, the repeated use of high-capacity models may pose practical limitations in terms of both time and resources.

Third, while our method improves factual accuracy and reduces hallucination, the validation pipeline is primarily designed for fact-based QA. This makes it less suitable for tasks involving subjective reasoning, commonsense inference, or open-ended discussion questions. Extending the pipeline to handle such cases would require fundamentally different validation strategies that go beyond factual grounding.

Finally, since the student models are trained on data generated by large models (used in both synthesis and supervision), there is a risk of knowledge leakage or model bias propagation. The synthetic data may overrepresent patterns and linguistic preferences from the teacher models, potentially limiting the generalizability of the student models trained on it.

Ethics Statement

Potential Risks The primary risk associated with this work lies in the possibility of propagating factual inaccuracies or biases through automatically synthesized data. While our validation pipeline aims to minimize hallucinations and ensure factual correctness, it may not catch all subtle errors. Additionally, overreliance on large language models for data generation could inadvertently reinforce biases encoded in those models. Our approach does not involve any sensitive personal data or downstream applications that could directly harm individuals.

Annotator Recruitment To verify and refine the samples in our *M*^{2}*QA-Bench* dataset, we recruited three human annotators with prior experience in NLP and data annotation (two men and one woman). These annotators were compensated fairly at a rate of $25 per hour to reflect their expertise and time investment. All annotators were provided with detailed task descriptions and underwent an informed consent process prior to participation. The annotation process was conducted in accordance with ethical research guidelines and ensured voluntary participation and data confidentiality.

Evaluator Recruitment For the human evaluation component of our study, we recruited twelve evaluators to compare answers generated with and without our validation pipeline. Evaluators were compensated at a rate of $10 per hour and participated voluntarily after giving informed consent. They were clearly informed about the nature and purpose of the task. We ensured the task was low-risk, did not involve sensitive content, and that participation remained anonymous and non-intrusive.

Consent and Data Privacy All participants in both annotation and evaluation roles were briefed on the nature of the task and explicitly consented to take part in the study. No personally identifiable information was collected or stored during any part of the research process. All data generated and reviewed by annotators and evaluators remained anonymous and was used strictly for academic research purposes.

Use of AI Assistants We used AI assistants such as GitHub Copilot and ChatGPT to support coding, text editing, and formatting tasks during the development of this paper and the implementation of our framework. These tools were employed to accelerate workflow and refine writing, but all conceptual, experimental, and analytical decisions were made by the authors. We ensured that no sensitive data was provided to these tools during usage.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, and Harkirat et al. Behl. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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

Abhinav Jauhri Abhimanyu Dubey, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, and Artem Korenev et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yinghsan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. WebQA: Multihop and Multimodal QA.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Mingda Chen, Xilun Chen, and Wen-tau Yih. 2024b. Few-shot data synthesis for open domain multi-hop question answering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 190–208, St. Julian's, Malta. Association for Computational Linguistics.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023a. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Felipe González-Pizarro and Giuseppe Carenini. 2024. Neural multimodal topic modeling: A comprehensive evaluation. *arXiv preprint arXiv:2403.17308*.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Zhi Hong, Aswathy Ajith, James Pauloski, Eamon Duede, Kyle Chard, and Ian Foster. 2023. The diminishing returns of masked language models to science. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1270–1283, Toronto, Canada. Association for Computational Linguistics.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *Preprint*, arXiv:2403.12895.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Botian Jiang, Lei Li, Xiaonan Li, Zhaowei Li, Xiachong Feng, Lingpeng Kong, Qi Liu, and Xipeng Qiu. 2024. Understanding the role of llms in multimodal evaluation benchmarks. *arXiv preprint arXiv:2410.12329*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024a. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *Preprint*, arXiv:2310.07849.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.

Shivam Mehta, Anna Deichler, Jim O'regan, Birger Moëll, Jonas Beskow, Gustav Eje Henter, and Simon Alexanderson. 2024. Fake it to make it: Using synthetic data to remedy the data shortage in joint multimodal speech-and-gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1964.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, and Jeff Belgum et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Atula Tejaswi, Yoonsang Lee, Sujay Sanghavi, and Eunsol Choi. 2024. Rare: Retrieval augmented retrieval with in-context examples. *Preprint*, arXiv:2410.20088.

Ken Tsui. 2024. AnyClassifier.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakazad, Tongshuang Wu, and Graham Neubig. 2024. Synthetic multimodal question generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12960–12993, Miami, Florida, USA. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. *Preprint*, arXiv:1905.08284.

Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang. 2023. Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5223–5234, New York, NY, USA. Association for Computing Machinery.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Ying Shan, and Yansong Tang. 2024. VoCo-LLaMA: Towards Vision Compression with Large Language Models. *arXiv preprint arXiv:2406.12275*.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou,

Mingqian He, Yanna Ma, Weiming Lu, and 1 others. 2024. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. *arXiv preprint arXiv:2407.07053*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

A Prompts

In our data generation pipeline, FM^2DS, which incorporates LVLMs, we carefully designed prompts to guide the model through tasks involving cross-modal reasoning and data synthesis. Each prompt was carefully designed with specific elements to ensure precision, clarity, and completeness in achieving the task's objectives, while also minimizing the need for error correction during the evaluation process. In the following sections, we outline the rationale behind the structure and components of these prompts.

A.1 Question Generation

Prompt 1 show the prompt used for question generation. Using this prompt, we ask the model to create multi-hop questions that require information from all provided documents and modalities (e.g., text and images) to answer. The key aim is to design questions that are unanswerable if any one document or one modality is given, promoting the need for multi-document and multimodal reasoning. It ensures the model generates questions that require synthesizing information from diverse sources to form a comprehensive understanding. To avoid duplicate data generation, if the generated question was already present in the dataset, we reused the same prompt but included the previously generated questions from the same set of documents. The model was then instructed to generate a new, unique question.

A.2 Answer Generation

The prompt for answer generation directs the model to analyze multiple documents, encompassing both text and images, to address the given question. It emphasizes integrating and synthesizing information from all sources to deliver the most accurate and comprehensive response. The prompt ensures that the model considers all modalities and documents without relying solely on a single source or the model's pre-trained knowledge, focusing exclusively on the provided materials. Refer to Prompt 2 for the answer-generation prompt.

Question Generation Prompt

```
Generate a multi-hop question based on the provided information. A multi-hop question requires the model to utilize information from all available documents in combination to reach the correct answer. Specifically, the question should be designed to be unanswerable if any one of the documents is missing. Furthermore, focus on creating questions that compel the model to extract and synthesize relevant information across multiple modalities—such as images and text. This means that answering the question correctly will demand integrating insights from each source and modality, making it impossible to arrive at an accurate answer using any single document or modality alone.

Here are the documents:
[Documents]

Here are examples:
[Example(s)]
```

Prompt 1: The prompt for question generation defines what constitutes a multi-hop question and instructs the model to create a multimodal and multihop question based on the provided documents. It emphasizes that the question should require information from multiple modalities and multiple given documents to be answered, similar to the given example(s).

A.3 Query Generation

As illustrated in Prompt 3, in query generation, the model is tasked with explaining the step-by-step process used to extract relevant information from the documents and determine the answer based on the extracted snippets. This task emphasizes transparency by requiring the model to identify the relevant sections of each document and describe how information from multiple sources is retrieved and combined to arrive at the correct answer, promoting explainability in the model's reasoning process.

Answer Generation Prompt

```
You are provided with multiple documents, including both textual content and images, along with a question. Your task is to carefully review each document, analyze the images, and derive an answer based on the information contained across all sources. Aim to combine insights from both documents and across modalities to deliver the most accurate and comprehensive response possible.

Question:
[Question]

Here are the documents:
[Documents]

Here are examples:
[Example(s)]
```

Prompt 2: The answer generation prompt instructs the model to answer the question solely based on the provided documents, utilizing all available modalities, without relying on its pre-trained knowledge.

Query Generation Prompt

```
You are provided with multiple documents, a question, and the answer. Your task is to explain the step-by-step process you would use to extract and verify the answer using information from the documents and various modalities. Clearly identify each document title and relevant sections, and describe how you locate, interpret, and integrate information across both documents to derive the correct answer.

Question:
[Question]

Answer:
[Answer]

Here are the documents:
[Documents]
```

Prompt 3: The query generation prompt instructs the model to provide a step-by-step plan for extracting relevant information needed to answer the question.

B Experimental Settings

In this work, we conducted experiments on a cluster of **8 NVIDIA H100 80GB GPUs**. The distributed setup allowed us to efficiently scale our fine-tuning process across multiple devices. The fine-tuning process was carried out using low-rank adaptation (LoRA) (Hu et al., 2022), a technique for efficient adaptation of pretrained models with low-rank matrices, reducing the number of trainable parameters. The key hyperparameters used in the fine-tuning procedure include a learning rate of 1e-4, a batch size of 8 per device (totaling 64 across 8 devices), LoRA rank set to 8, LoRA alpha set to 32, a weight decay of 0.01, and the number of epochs was 5. Additionally, AdamW optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e{-}8$. The models were fine-tuned using mixed-precision training to take full advantage of the 80GB memory on each H100 GPU. For inference time, we set the temperature to 0.7, which strikes a balance between randomness and coherence in the model's responses, producing more varied outputs without sacrificing too much quality. This setup ensured efficient usage of computational resources while maintaining high model performance.

C The Effect of the Number of In-Context Examples

Number of Shots	MultiModalQA		WebQA	
	EM	F1	EM	F1
zero-shot	66.42	75.1	71.05	79.3
one-shot	73.92	84.27	81.27	90.21
two-shot	74.3	83.71	81.5	91.37
three-shot	74.45	85.39	81.6	91.21

Table 6: Effect of the number of in-context documents on the performance of Intervl-2-8B on MultiModalQA and WebQA datasets.

Table 6 presents the results of evaluating the Intervl-2-8B model with varying numbers of in-context examples on the MultiModalQA and WebQA datasets. In the zero-shot setting, *FM*^{2}*DS* exhibits limited understanding of multimodal multi-hop question answering, and occasionally circumvents the validation step by simply generating a

question that is not multihop. For example, "*Looking at the image of the Eiffel Tower, what engineering innovation allowed it to surpass previous structures in height?*" prompts the model to use the image, but the answer is available in the page's text on tall structures. As we move from zero-shot to one-shot, there is a significant boost in EM and F1 scores, reflecting improved performance with minimal context. The improvement from one-shot to two-shot is marginal, suggesting diminishing returns. With three in-context samples, the gains become minimal, indicating that additional samples beyond two provide little benefit. This diminishing return may stem from the model's limited context window, which restricts its ability to fully utilize large in-context samples (Kaplan et al., 2020; Bertsch et al., 2024).

D Comparing Methods For Data Synthesis

To evaluate the effectiveness of different methods for synthetic data generation, we compared three prominent language models: GPT-4o, Claude 3.5 Sonnet, and Llama-3.2-90B, as shown in Table 7. Using Intervl-2-8B with 5K fine-tuning samples as our baseline model, we tested the quality of generated data on two distinct datasets: MultiModalQA and WebQA. The results, measured using EM and F1 scores, demonstrate that GPT-4o consistently outperforms other models across both datasets. We also see that Llama-3.2-90B shows competitive performance as an open-source model with less number of parameters, particularly in WebQA tasks. Claude 3.5 Sonnet generally yields lower scores across both datasets. As shown in Figure 4, Claude-3.5-Sonnet outperforms Llama-3.2-90B, which may be attributed to differences in the tasks that were included during their respective training phases. This observation can be further investigated.

Model	MultiModalQA		WebQA	
	EM	F1	EM	F1
GPT-4o	73.92	84.27	81.27	90.21
Claude-3.5-Sonnet	68.35	76.01	76.98	83.86
Llama-3.2-90B	70.64	76.32	79.81	86.75

Table 7: Performance comparison of different models for data generation on test datasets.

E Investigating *FM^{2}DS* on Other Models

In addition to the models discussed in Section 5, we explored other model families, including Idefics3 (Laurençon et al., 2024a), mPLUG-DocOwl-1.5 (Hu et al., 2024), and Phi-3.5-Vision-Instruct (Abdin et al., 2024a), as well as larger versions within the explored families presented in Table 3. The results in Table 8 demonstrate the reliability of our data synthesis approach, which consistently enhances model performance across all models and sizes compared to an equivalent number of real samples.

As Table 8 shows, within the same model architecture, as the number of parameters increases and the model complexity grows (e.g., InternVL-2), the performance generally improves, including the pre-trained version. These models also exhibit more effective learning, especially when provided with synthesized data generated by *FM^{2}DS*, which makes the learning process more efficient. Moreover, Idefics-3 shows notable improvement over its predecessor, Idefics-2, indicating that the newer version has a better visual reasoning. When comparing mPLUG-DocOwl-1.5 with models like InternVL-2, Idefics-2, and Idefics-3, it demonstrates relatively lower performance. This could be attributed to the training objective of mPLUG-DocOwl-1.5, which focuses on multi-grained text recognition and parsing, potentially resulting in weaker performance when visual reasoning is required. Nevertheless, this model still outperforms LLaVA-1.6-7B overall, which might be due to the simpler structure of the LLaVA-1.6 family. Finally, Phi-3.5-Vision-Instruct, despite having fewer parameters compared to other models, performs competitively with other models and surpasses LLaVA-1.6-7B in performance.

F Human Evaluation Details

To facilitate a rigorous human evaluation of our answer validation component, we created a Google Form to recruit participants willing to contribute to our evaluation. We shared this form widely and will acknowledge the contributions of participating individuals in the acknowledgment section of the paper's camera-ready version.

After registration, participants were divided into four batches (three participants per batch, each assigned 25 samples, 100 in total) and given access to a custom evaluation app, shown in Figure 6, to review the samples in their assigned batch. This

Model	Test Dataset									
	MultiModalQA					WebQA				
		EM		F1			EM		F1	
	FT (Real/Syn)	Real	Syn	Real	Syn	FT (Real/Syn)	Real	Syn	Real	Syn
LLaVa-1.6-34B	10k/10k	79.41	**80.92**	82.55	**83.71**	10k/10k	84.41	**84.94**	85.58	**85.79**
LLaVa-1.6-34B	23.8k/16k	84.83	**85.29**	85.51	**86.42**	34.2k/13k	86.48	**87.49**	**88.18**	**88.18**
InternVL-2-40B	10k/10k	82.18	**83.56**	89.24	**90.2**	10k/10k	87.67	**89.77**	92.76	**93.82**
InternVL-2-40B	23.8k/15k	86.63	**87.27**	**91.73**	91.44	34.2k/14k	89.77	**90.32**	**93.19**	**93.19**
InternVL-2-76B	10k/10k	83.95	**86.15**	89.76	**90.72**	10k/10k	88.12	**90.32**	93.19	**94.77**
InternVL-2-76B	23.8k/14k	90.82	**91.34**	92.79	**93.81**	34.2k/14k	93.14	**93.65**	**94.06**	93.82
Idefics-3-8B	10k/10k	76.42	**77.56**	86.74	**88.23**	10k/10k	82.48	**84.62**	89.12	**90.09**
Idefics-3-8B	23.8k/19k	82.18	**83.56**	90.27	**91.81**	34.2k/15k	86.49	**87.77**	92.55	**93.31**
Phi-3.5-Vision-Instruct-4.2B	10k/10k	69.43	**70.25**	75.35	**77.57**	10k/10k	78.31	**80.22**	84.5	**85.18**
Phi-3.5-Vision-Instruct-4.2B	23.8k/22k	77.85	**78.79**	82.59	**84.61**	34.2k/19k	80.11	**81.27**	85.34	**87.48**
mPLUG-DocOwl-1.5-8B	10k/10k	72.24	**74.82**	78.82	**79.49**	10k/10k	81.27	**83.86**	87.52	**88.75**
mPLUG-DocOwl-1.5-8B	23.8k/20k	79.41	**80.12**	84.69	**87.07**	34.2k/17k	82.48	**84.42**	87.93	**89.21**
LLaVa-1.6-34B	None	60.11		64.06		None	64.33		70.82	
InternVL-2-40B	None	72.93		77.42		None	76.98		82.33	
InternVL-2-76B	None	**75.32**		**79.32**		None	**78.31**		**85.41**	
Idefics-3-8B	None	61.27		69.74		None	69.88		76.39	
Phi-3.5-Vision-Instruct-4.2B	None	55.78		62.16		None	63.68		69.74	
mPLUG-DocOwl-1.5-8B	None	58.46		64.02		None	66.38		71.26	

Table 8: Comparison of model performance across various architectures, sizes, and sample sources (real vs. synthesized by FM^2DS). The models were evaluated on 10k samples and the full dataset (23.8k samples for MultiModalQA and 34.2k samples for WebQA). When comparing models tuned on synthesized data with those trained on the full training set, the smallest number of synthetic samples (divisible by 1k) that outperforms models trained on the full datasets is reported. For real sample evaluations, the WebQA training set is used for testing on the WebQA test set, and the same applies to MultiModalQA. Models trained with synthesized samples consistently outperform those trained with equivalent numbers of real samples.

application was designed to streamline the evaluation process and ensure consistency across participants. For each question, participants could review the question text, the associated Wikipedia pages, and the generated answers from two methods—one method utilizing the answer validation component and the other without it. To minimize user bias, the application randomly alternated the positioning of the methods' answers (labeling them as "Answer A" and "Answer B") so that users could not develop a tendency to select one model over the other based on position alone. After examining the question and relevant Wikipedia content, users were asked to select one of four options to indicate their assessment of answer accuracy: (1) Answer A is correct, (2) Answer B is correct, (3) both answers are correct, or (4) neither answer is correct.

In addition to these selections, participants had the option to provide a brief rationale for their choices. Although they have not been investigated for this research, these optional feedbacks were encouraged, as they offer valuable insights for qualitative analysis and potential future improvements in answer validation accuracy. The combination of structured and open-ended responses enhances the robustness of our evaluation and offers a more comprehensive view of user judgments, which we may explore in future iterations of our data synthesis methodology.

The evaluators had diverse academic and professional backgrounds, including graduate students in computer science, data science researchers, and software engineers with experience in NLP and machine learning. All evaluators were proficient in English and had prior familiarity with Wikipedia-style content and fact-based question answering tasks. This diversity contributed to reliable judgment across a wide range of topics and ensured that participants had the necessary background to assess factual correctness and relevance accurately. In total, twelve individuals participated in the evaluation: seven men and five women.

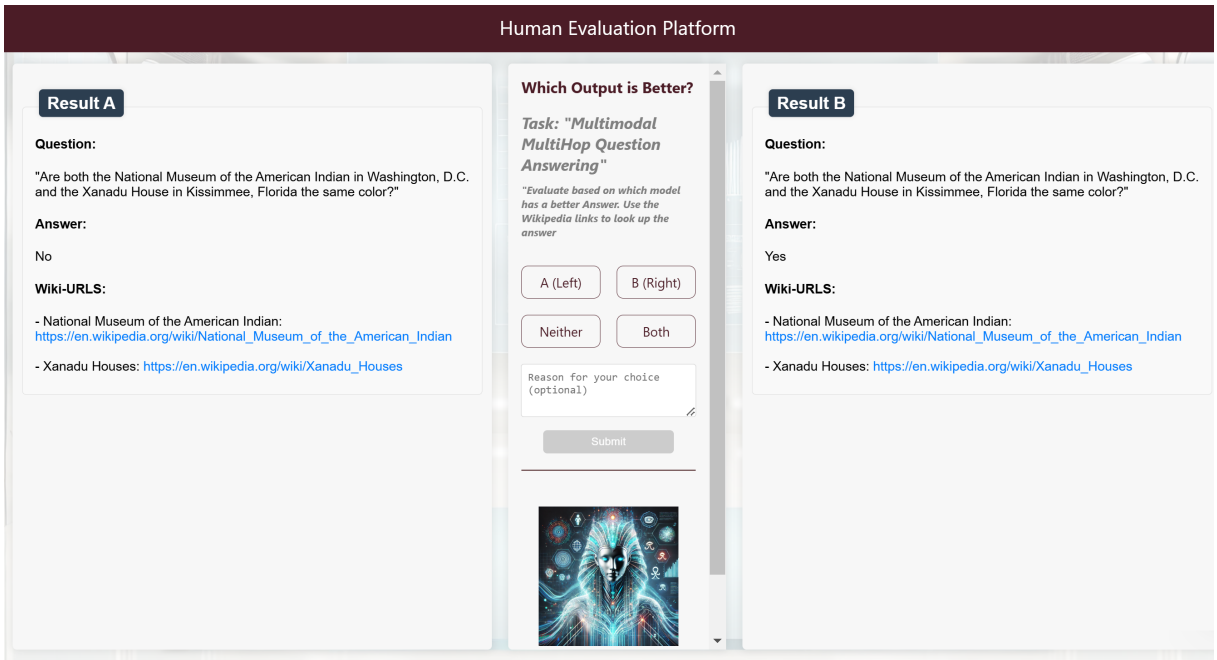

Figure 6: The custom evaluation application was used for human evaluation. The application presents each participant with a randomly selected question, relevant Wikipedia pages, and two model-generated answers labelled as *Answer A* and *Answer B*. One answer is generated by the pipeline with validation, while the other comes from the pipeline without it. Participants are asked to choose the correct answer and optionally provide feedback on their choice. To minimize bias, the application randomizes the position of each model's answer.

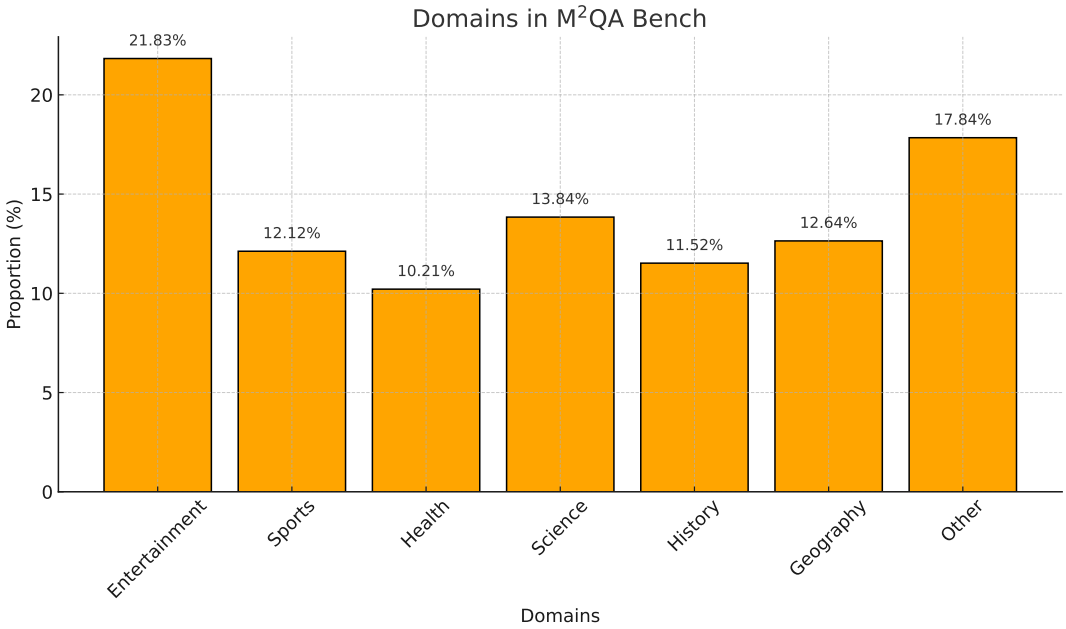

Figure 7: Distribution of domains in *M²QA-Bench*.

G Additional Statistics & Information on *M²QA-Bench*

As illustrated in Figure 7, *M²QA-Bench* encompasses a diverse range of domains. Additionally, the answers span various types of named entities, including people, products, works of art, and more. Figure 8 presents the distribution of named entities found in the answers.

To further examine the diversity of questions in our benchmark—which also reflects the overall characteristics of the data generated by *FM²DS*—we conducted a 2D t-SNE analysis of

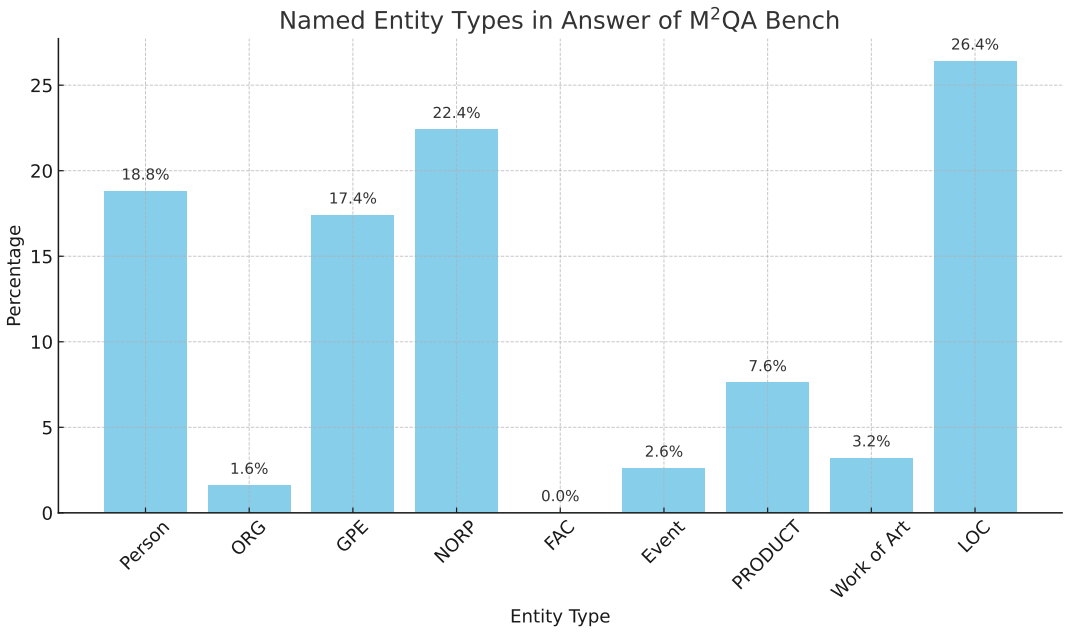

Figure 8: Distribution of named entities in answers in *M²QA-Bench*.

question embeddings using ModernBERT (Warner et al., 2024). We sampled 500 questions each from *M²QA-Bench*, MMQA, WebQA, and ScienceQA (Lu et al., 2022b). ScienceQA serves as a fully human-authored dataset, while MMQA and WebQA primarily use templated questions. As shown in Figure 9, MMQA and WebQA display the least diversity. In contrast, *M²QA-Bench*, which includes questions generated from *FM²DS*, demonstrates greater similarity to human-generated data, reflecting a reduced domain gap and improved diversity compared to MMQA and WebQA.

H *FM²DS* Generalizability to Other Domains

FM²DS can generate domain-specific synthetic data using just a single in-context example and user-provided documents, enabling even small LVLMs to handle domain-specific multimodal multihop QA. As shown in Figure 7, *FM²DS*'s data (including *M²QA-Bench*) spans various domains. To further assess its generalizability, we trained InternVL-2-8B on 5k synthesized samples and evaluated it on the health-related dataset **PMC-VQA** (Zhang et al., 2023) using 50k test samples. We then compared its performance to the pretrained InternVL-2-8B on the same test set. Training improved accuracy from **52.82** to **65.72**, demonstrating the benefits of our synthetic data.

I *M²QA-Bench* Examples

FM²DS uses LVLMs to generate multimodal and multihop questions based on the given documents and evaluate their answers. These samples aim to emulate few-shot examples typically provided to guide the model's behavior in a structured and relevant manner.

In some cases, the questions focus on understanding facts from different modalities—such as images, text, and tables—within the grouped documents and finding the answer from one of them. For example, in the case of the question shown in Figure 10:

> *How many people died in the event shown in the photograph "Raising the Flag on Iwo Jima" from the country shown in the picture?*

LVLM is tasked with combining information from two documents: *Raising the Flag on Iwo Jima* and *Battle of Iwo Jima*. Here, the hyperlink between the two documents served as the connection between two docments. The model identifies that the photograph depicts American soldiers (based on the USA flag) and cross-references the table from the *Battle of Iwo Jima* document to determine that 539 people from the USA were killed. This demonstrates how the model synthesizes information across modalities to form an accurate response. Afterward, the model generates queries, serving

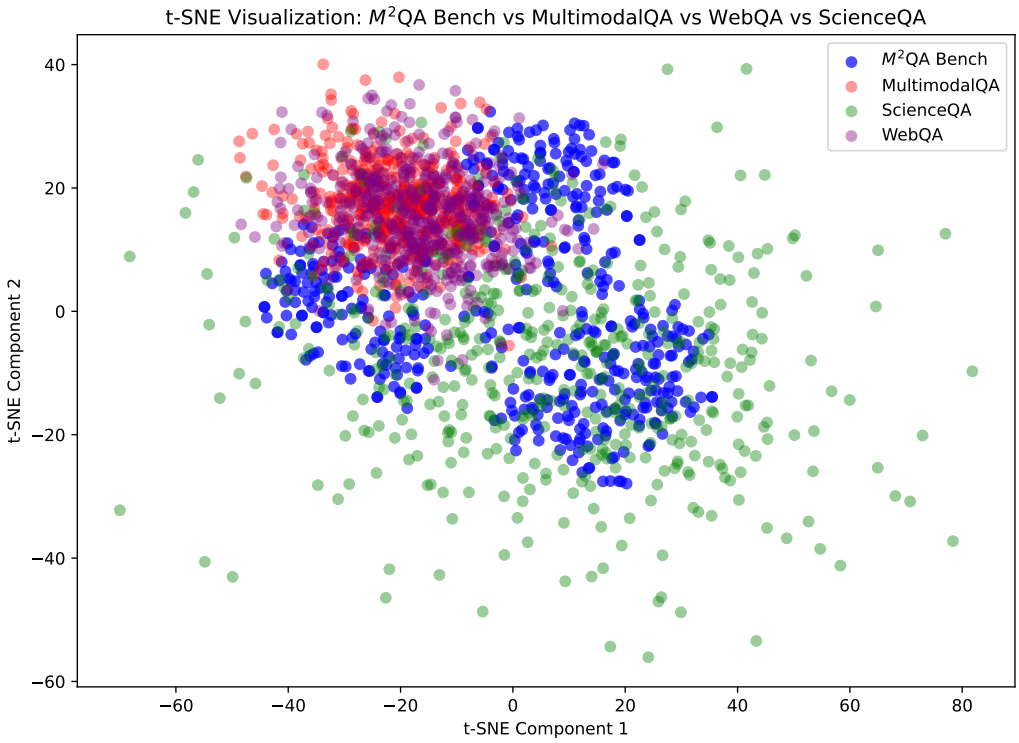

Figure 9: 2D t-sne visualization of ModernBERT embeddings of questions between *M^2QA-Bench*, WebQA, MultimodalQA, and ScienceQA.

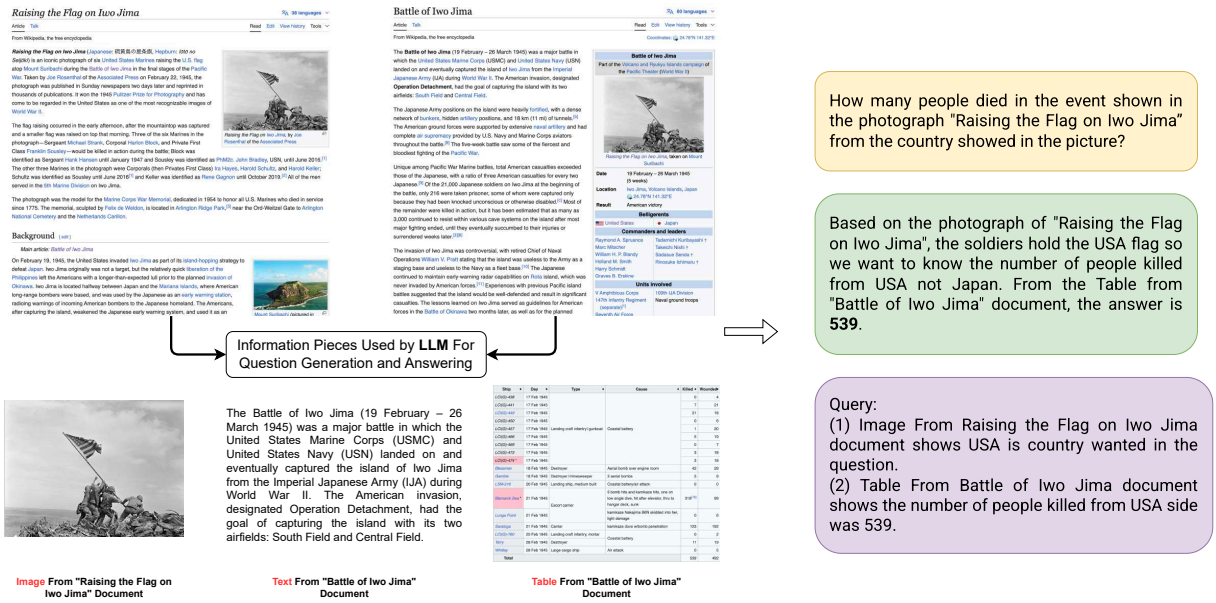

Figure 10: Multimodal and multihop reasoning example from *M^2QA-Bench* where the model answers a question about the photograph "Raising the Flag on Iwo Jima" by synthesizing information from linked documents through a hyperlink, leveraging both visual and tabular data to determine the number of casualties from the USA.

as a step-by-step guide to extract relevant information from the documents. Using the extracted snippets, it then answers the question. For instance, the model would need to locate the image *Raising the Flag on Iwo Jima* to determine the country mentioned in the question, which is the USA. Next,

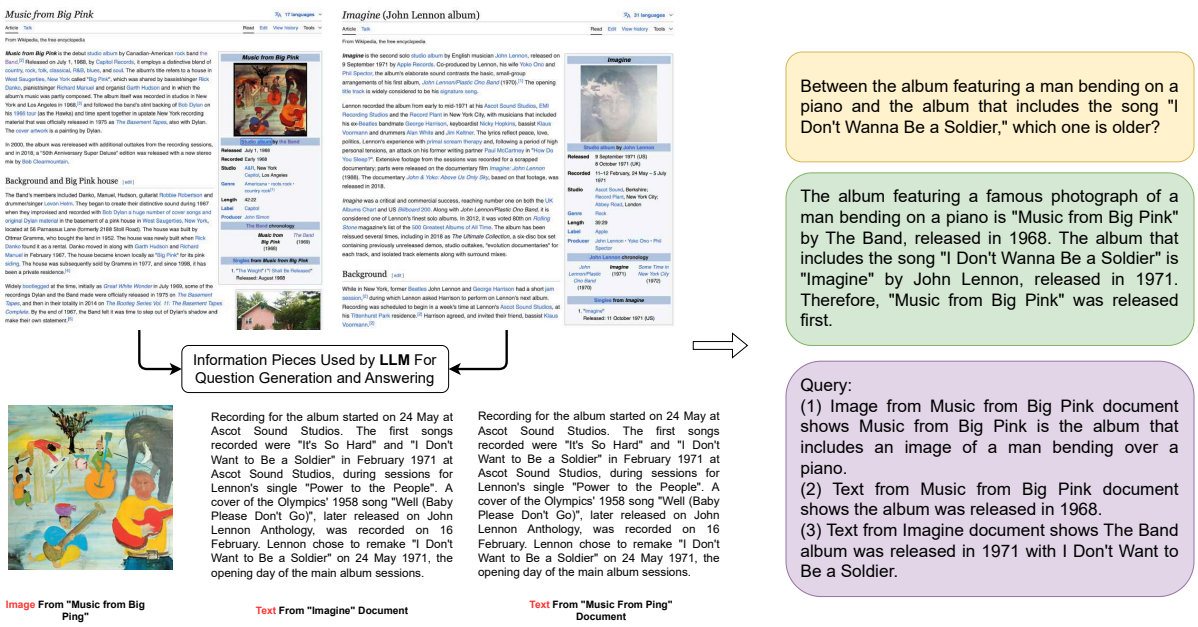

Figure 11: Multimodal multihop reasoning example from M^2QA-*Bench* where the model compares the release dates of two albums, "Music from Big Pink" and "Imagine," using textual and visual cues. The documents are connected through their shared topic, "music," and the answer is determined as the title of the earlier-released album.

by referencing the table in the *Battle of Iwo Jima* document, it provides the final answer.

In other cases, the questions involve comparing elements between objects in two different documents, where the answer is typically the title of one of the documents provided. For example, the question shown in Figure 11:

> *Which album was released first: the one featuring a famous photograph of a man bending on a piano or the album that includes the song "I Don't Wanna Be a Soldier"?*

requires the model to compare temporal information across two documents: *Music from Big Pink* and *Imagine*. The model identifies that *Music from Big Pink*, featuring a photograph of a man bending on a piano, was released in 1968, while *Imagine*, containing the song "I Don't Wanna Be a Soldier," was released in 1971. Therefore, the answer is *Music from Big Pink*. In this case, the documents were connected through their shared topic, music. The query generation in this example is similar to the first but differs slightly, as three information snippets are key to answering the question, making the query three steps long.

J Qualitative Analysis

In the qualitative analysis, we compared three critical factors influencing model responses: model architecture, fine-tuning (FT) dataset (real samples or synthesized samples), and model size. To examine the effects of model architecture and FT dataset, we used InternVL-2-8B, LLaVA-1.6-7B, and Idefics-2-8B, fine-tuning them on both real and synthetic data generated by FM^2DS. For analyzing the impact of model size, all versions of InternVL-2 were trained on the synthetic data. All of the mentioned models were fine-tuned on 5k samples.

This analysis was conducted for 100 samples from each of the following benchmarks: (1) M^2QA-*Bench*, (2) MultiModalQA, and (3) WebQA. The results are presented in Tables 9, 10, and 11. The responses generated by different models were analyzed across these datasets, focusing on the following metrics:

1. Model accuracy using the exact match (EM) metric.
2. Hallucination rate, corresponding to instances where the model generated wrong answer based on its pre-trained knowledge instead of the provided document.
3. Model accuracy with EM metric for samples including image modality (may include other modalities).
4. Model accuracy with EM metric for samples including table modality (may include other modalities).

5. Model accuracy with EM metric for samples including both image and table modalities.

For WebQA, which only incorporates text and image modalities, the last three metrics were not applicable. Additionally, the distribution of modalities across samples for MultiModalQA and *M²QA-Bench* was as follows:

- ***M²QA-Bench*:** 66 samples included image modality, 62 samples included table modality, and 28 samples included both image and table modalities.

- **MultiModalQA:** 61 samples included image modality, 54 samples included table modality, and 15 samples included both image and table modalities.

Overall, in all benchmarks, model hallucination rates decreased as model complexity and parameter count increased, resulting in more accurate answers across all modalities (e.g., see Figure 15 for an example output of these models). Larger models consistently outperformed smaller models on both modalities. Regarding synthetic data, fine-tuning on data generated by *FM²DS* significantly reduced hallucination and improved performance across all modalities. While the hallucination rates among different model families are relatively similar, all models occasionally generate answers based on their pre-trained knowledge rather than the provided document. Fine-tuning on data generated by *FM²DS* effectively alleviates this issue. Among the models, as shown in Table 3, LLaVA-1.6 exhibited the poorest performance and the highest likelihood of hallucination, followed by Idefics-2, with InternVL-2 demonstrating the best performance.

Regarding the effect of modalities, results from Tables 9 and 10 suggest that the modalities themselves are not the most critical factor. Instead, the complexity of how the question integrates the modalities plays a more significant role. For *M²QA-Bench*, models performed better when visual understanding was not required, with tables and text being the primary contributors to the results. In contrast, for MultiModalQA, models tended to perform better on image-based questions, highlighting the importance of how the question leverages the modalities. For questions involving both modalities, smaller models struggled more to produce correct answers, while larger models performed better in terms of EM. It is important to note, however, that due to the substantial difference in the number of samples containing both image and table modalities compared to those with only one modality, the reported results are not directly comparable. Refer to Figures 13 and 14 for the outputs of different model families fine-tuned on either real or synthesized data. Moreover, Figure 15 shows outputs from different model sizes within the same family, fine-tuned on either real or synthesized data.

K Synthesizing Data vs. Paraphrasing Exisiting Human Annotated Datasets

Paraphrasing questions from existing datasets introduces surface-level linguistic changes but preserves the original semantic intent and reasoning pathways, offering only marginal improvements in model training. In contrast, the data synthesized by *FM²DS*is intentionally crafted to introduce diverse question structures, span multiple domains, and require varied types of reasoning, pushing models toward more comprehensive multimodal understanding. To compare these approaches, we trained InternVL-2-8B using 1k samples from three settings: (I) the original MultimodalQA dataset, (II) a paraphrased version of MultimodalQA where questions were reworded using GPT-4o with the prompt "Please paraphrase the following question: [Question]" and (III) synthesized samples generated by *FM²DS*. For all conditions, we used full Wikipedia documents as sources. Figure 12 presents the results, showing that while paraphrasing provides a slight improvement, synthesizing new, high-quality samples with *FM²DS* leads to a substantial performance gain.

L Statistics on Usages of Each Validation Stage of *FM²DS*

As described in Section 3 and illustrated in Figure 2, *FM²DS* incorporates multiple validation stages to enhance data quality. It is essential to analyze how frequently each stage rejects the initially generated outputs. Table 12 presents statistics based on generating 1,000 examples using GPT-4o. Among the stages, question validation has the highest rejection rate, suggesting that this step is the most challenging. This may be because generating a question requires the model to synthesize all relevant knowledge and fully grasp the context. In contrast, answer validation benefits from the guidance provided by the question, making the task relatively easier. Query validation appears to be even more

Model	Trained On	EM ↑	Hallucination ↓	EM (Table)	EM (Image)	EM (Image&Table)
InternVL-2-8B	Real	0.43	0.67	0.56	0.51	0.46
InternVL-2-8B	Synth	**0.57**	**0.51**	**0.77**	**0.68**	**0.64**
Idefics-2-8B	Real	0.39	0.72	0.54	0.44	0.43
Idefics-2-8B	Synth	**0.55**	**0.68**	**0.71**	**0.62**	**0.53**
LLaVA-1.6-7B	Real	0.35	0.71	0.43	0.39	0.32
LLaVA-1.6-7B	Synth	**0.47**	**0.51**	**0.61**	**0.46**	**0.39**
InternVL-2-26B	Synth	0.61	0.54	0.87	0.74	0.75
InternVL-2-40B	Synth	0.64	0.5	0.89	0.8	0.78
InternVL-2-76B	Synth	**0.72**	**0.28**	**0.98**	**0.95**	**0.92**

Table 9: Performance comparison of different model families fine-tuned on real and synthesized data on *M^2QA-Bench*. The ratios for EM scores and hallucination were calculated from filtered data (e.g., hallucination as the proportion of hallucinated responses to incorrect answers). In the table, ↑ indicates that larger values are better (all EM values), while ↓ indicates that smaller values are better (hallucination rate). The scores for EM (Table) and EM (Image) may include samples that also contain other modalities. Larger models and those fine-tuned on synthesized data generally show improved performance with reduced hallucination rates. In *M^2QA-Bench*, models demonstrate higher performance on questions involving table modality compared to those involving images.

Model	Trained On	EM ↑	Hallucination ↓	EM (Table)↑	EM (Image)↑	EM (Image & Table)↑
InternVL-2-8B	Real	0.66	0.62	0.7	0.72	0.53
InternVL-2-8B	Synth	**0.68**	**0.41**	**0.75**	**0.78**	**0.67**
Idefics-2-8B	Real	0.61	0.64	0.61	0.63	0.33
Idefics-2-8B	Synth	**0.64**	**0.53**	**0.67**	**0.69**	**0.47**
LLaVA-1.6-7B	Real	0.59	0.66	0.57	0.56	0.2
LLaVA-1.6-7B	Synth	**0.62**	**0.39**	**0.64**	**0.61**	**0.33**
InternVL-2-26B	Synth	0.7	0.37	0.79	0.85	0.8
InternVL-2-40B	Synth	0.74	0.35	0.85	0.89	0.87
InternVL-2-76B	Synth	**0.8**	**0.15**	**0.93**	**0.94**	**0.93**

Table 10: Performance comparison of different model families fine-tuned on real and synthesized data on MultimodalQA. The ratios for EM scores and hallucination were calculated from filtered data (EM(Table) refers to the EM score calculated on samples that include the table modality). In the table, ↑ indicates that larger values are better (all EM values), while ↓ indicates that smaller values are better (hallucination rate). The scores for EM (Table) and EM (Image) may include samples that also contain other modalities. Larger models and those fine-tuned on synthesized data generally exhibit improved performance with reduced hallucination rates. In MultiModalQA, models demonstrate higher performance on questions involving the image modality compared to those involving tables.

straightforward, as it primarily involves formatting the reasoning steps—something the model has effectively done during answer generation. Additionally, the use of question-specific image captions during answer generation likely contributes to a lower error rate by helping the model locate the correct information only text modality.

Model	Train On	EM ↑	Hallucination ↓
InternVL-2-8B	Real	0.76	0.41
InternVL-2-8B	Synth	**0.78**	**0.32**
Idefics-2-8B	Real	0.72	0.48
Idefics-2-8B	Synth	**0.75**	**0.36**
LLaVA-1.6-7B	Real	0.70	0.53
LLaVA-1.6-7B	Synth	**0.74**	**0.34**
InternVL-2-26B	Synth	0.81	0.21
InternVL-2-40B	Synth	0.82	0.11
InternVL-2-76B	Synth	**0.85**	**0**

Table 11: Overall performance of different model families fine-tuned on real and synthesized data on WebQA. Hallucination is the proportion of hallucinated responses to incorrect answers. In this table, ↑ indicates that larger values are better, while ↓ indicates that smaller values are better. Fine-tuning on synthesized data consistently reduces hallucination rates and improves EM scores across all models, with larger models achieving the best performance. Unlike *M²QA-Bench* and MultModalQA, WebQA only includes image and text modality, as a result no EM(Image) and EM(Table) are reported.

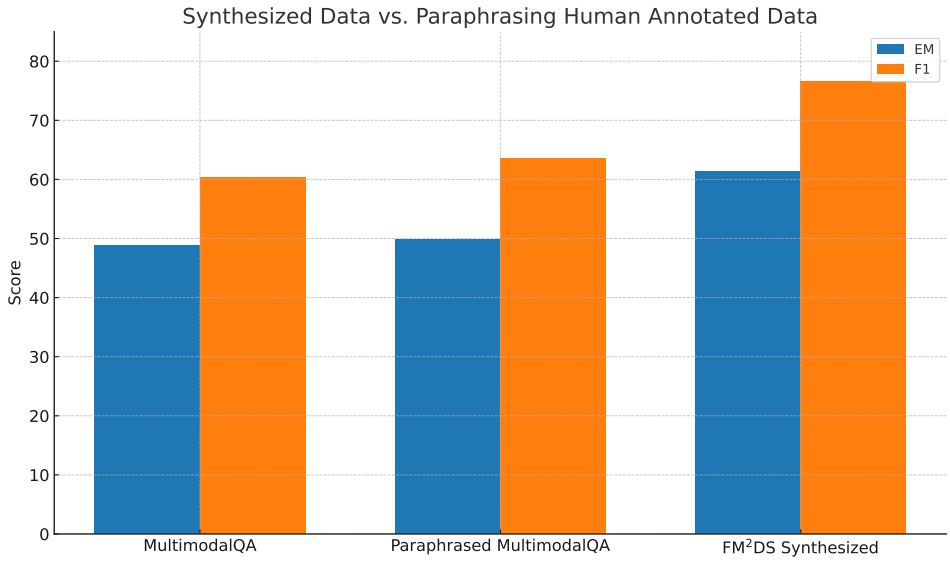

Figure 12: Performance comparison of InternVL-2-8B trained on 1k samples from three settings: original MultimodalQA, paraphrased MultimodalQA (reworded using GPT-4o), and fully synthesized data from *FM²DS*. While paraphrasing existing questions yields only modest gains, our synthesized samples lead to significantly higher performance, highlighting the value of generating diverse and structurally novel multihop multimodal questions.

Stage	Average Rejection Rate	Average Rejection Rate
Question Validation	131	11.58%
Answer Validation	76	7.06%
Query Validation	58	5.48%

Table 12: Key statistics of the proposed multimodal multihop question answering benchmark.

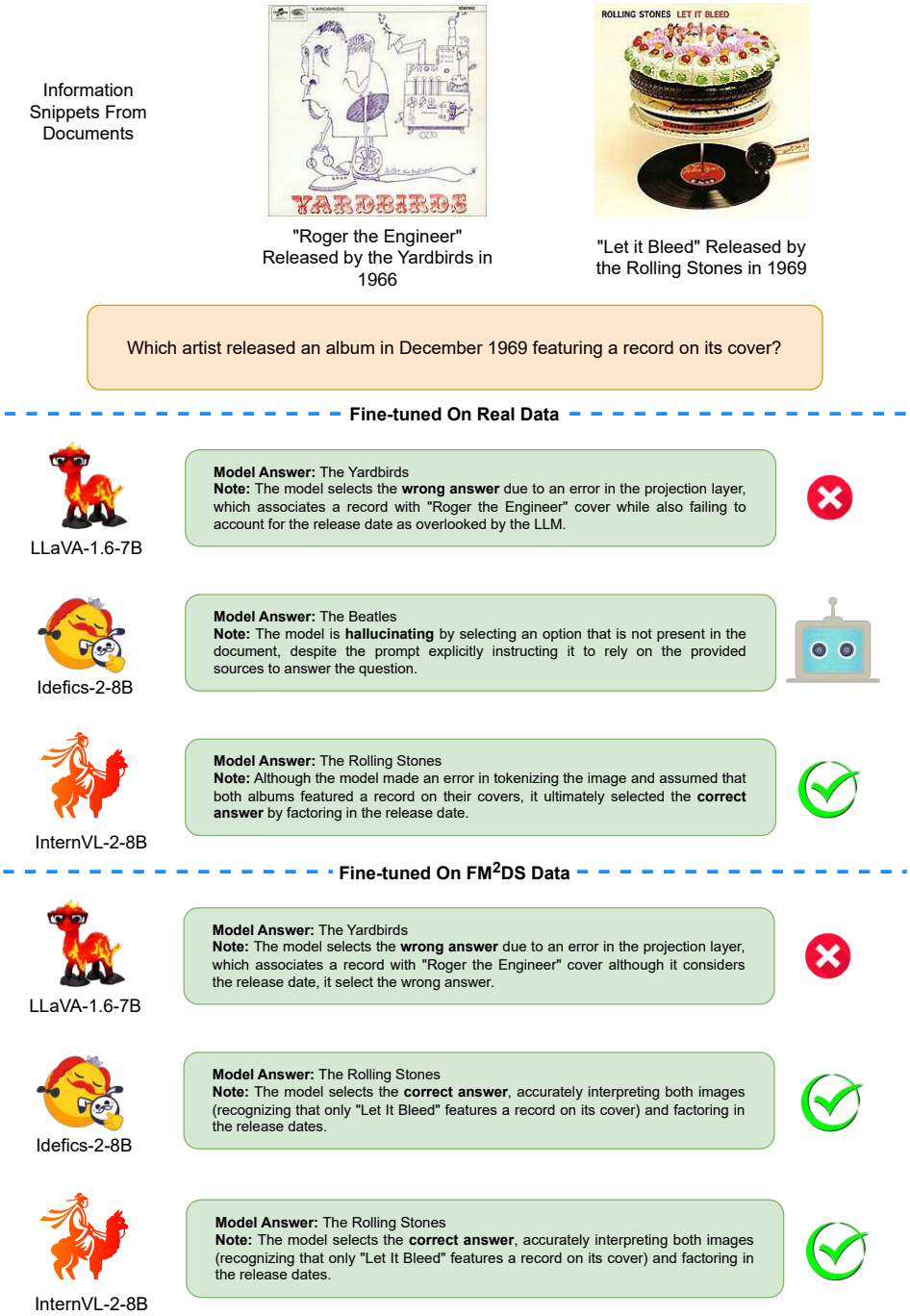

Figure 13: Analysis of model responses to the question: *"Which artist released an album in December 1969 featuring a record on its cover?"* from MultimodalQA dataset reveals that fine-tuning on *FM^{2}DS* eliminates hallucination (marked by the confused robot sign) seen in model fine-tuned on real data. This example highlights how fine-tuning improves reasoning by aligning the model's answers with both visual and textual evidence.

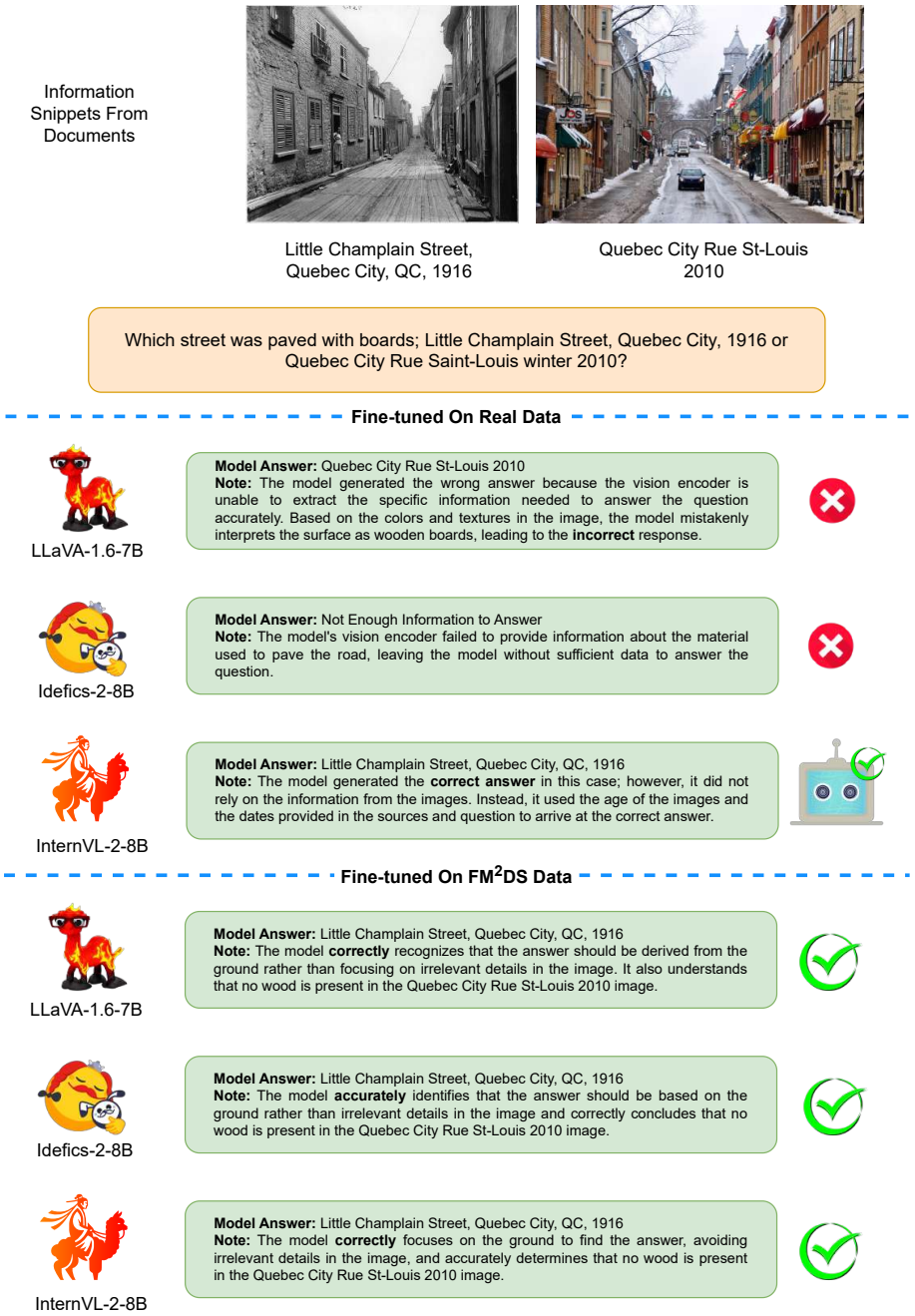

Figure 14: Analysis of model responses to the question: *"Which street was paved with boards; Little Champlain Street, Quebec City, 1916 or Quebec City Rue Saint-Louis winter 2010?"* from the WebQA dataset demonstrates that fine-tuning with *FM^{2}DS* data effectively eliminates hallucination (indicated by the confused robot sign). This example underscores fine-tuning with *FM^{2}DS*-generated data improves the model's focus on fine-grained visual details relevant to the question. Here, InternVL-2-8B fine-tuned on real data hallucinated but reached the correct answer using its pre-trained knowledge.

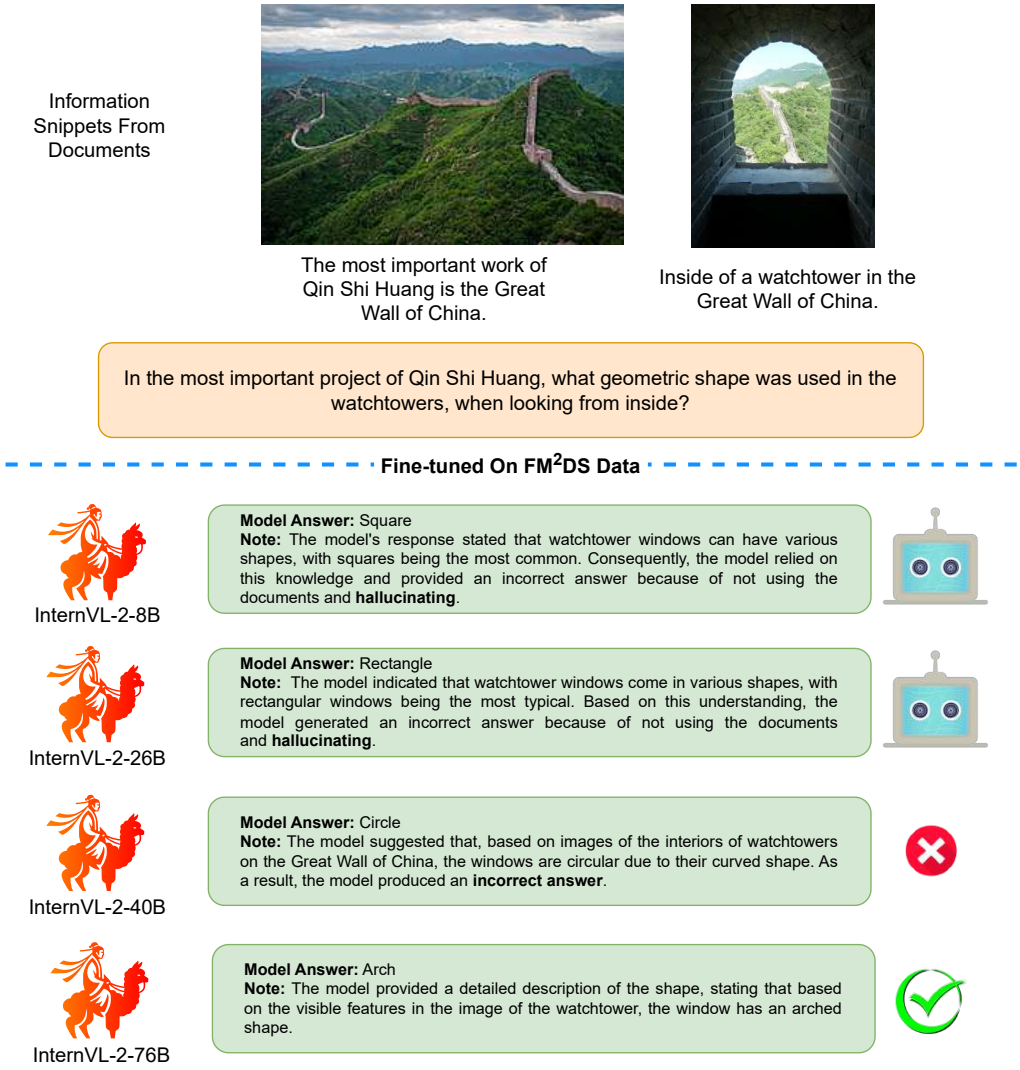

Figure 15: Responses from InternVL-2 models of various sizes (8B, 26B, 40B, and 76B) to the question: *"In the most important project of Qin Shi Huang, what geometric shape was used in the watchtowers when viewed from inside?"* from M^2*QA-Bench* illustrate that in examples like this, which requires detailed visual understanding, smaller models often hallucinate, providing inconsistent answers (e.g., square, rectangle) without grounding in the provided document. Larger models, however, perform better on this task and have less hallucination.