This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CRIT: Graph-Based Automatic Data Synthesis to Enhance Cross-Modal Multi-Hop Reasoning

Junyoung Sung[1] Seungwoo Lyu[1] Minjun Kim[1] Sumin An[1]
Arsha Nagrani[2] Paul Hongsuck Seo[1]
[1]Dept. of CSE, Korea University [2]Google DeepMind
{jys7451, dbtmddn41, ddomjun, suminan, phseo}@korea.ac.kr, arsha.nagrani@gmail.com

Abstract

Real-world reasoning often requires combining information across modalities, connecting textual context with visual cues in a multi-hop process. Yet, most multimodal benchmarks fail to capture this ability: they typically rely on single images or set of images, where answers can be inferred from a single modality alone. This limitation is mirrored in the training data, where interleaved image–text content rarely enforces complementary, multi-hop reasoning. As a result, Vision-Language Models (VLMs) frequently hallucinate and produce reasoning traces poorly grounded in visual evidence. To address this gap, we introduce CRIT, a new dataset and benchmark built with a graph-based automatic pipeline for generating complex cross-modal reasoning tasks. CRIT consists of diverse domains ranging from natural images, videos, and text-rich sources, and includes a manually verified test set for reliable evaluation. Experiments on this benchmark reveal that even state-of-the-art models struggle on such reasoning tasks. Models trained on CRIT show significant gains in cross-modal multi-hop reasoning, including strong improvements on SPIQA and other standard multimodal benchmarks.

1. Introduction

As humans, we are constantly interacting with a multimodal world. Real-world tasks often require cross-modal reasoning across multiple, interleaved sources of information. For instance, when following a DIY tutorial or a recipe illustrated with images, we may find ourselves continually cross-referencing textual instructions with a sequence of images, engaging in a multi-hop process to connect steps, tools, and outcomes [7, 11, 37]. This involves grounding textual descriptions (e.g., "fold the dough") to corresponding visual content (e.g., an image showing the dough being folded) and maintaining consistency of entities and states across modalities and steps.

Despite the importance of such tasks, most existing multimodal benchmarks do not adequately assess this capability. Current benchmarks, while valuable–typically present a single image [9, 35, 36, 38, 58], a single video, or a static set of images [14, 39, 53], with questions and answers represented in the textual domain. Even in datasets designed for multi-stage reasoning [40, 41, 59], the necessary information can often be inferred from a single modality alone, failing to test true cross-modal grounding. In Figure 1, we provide an example of the complex reasoning we target. The model must execute multi-hop, cross-modal reasoning to reach the correct answer. It must first identify the conceptual topic 'digital autonomy', which is associated with the student Elara Myles in the textual description, then determine from additional context that the 'laptop' symbolizes digital autonomy, and finally locate the laptop within the image to conclude that its color is 'silver'. Crucially, the multimodal information is *complementary*; the visual data provides attribute information while the text provides additional context unavailable in the pixels alone.

This gap in evaluation is mirrored by a lack of suitable post-training or supervised fine-tuning (SFT) data. While a lot of interleaved image-text data is used during training, it is unclear how much of it is truly complementary, and hence how much the model is forced to correlate the two modalities. Given the scarcity of such complex interleaved data, it is perhaps unsurprising that existing Vision-Language Models (VLMs) struggle on such complex reasoning tasks. As we show in our experiments, when prompted for step-by-step reasoning, Chain of Thought (CoT) traces output by models are often poorly grounded in the visual and textual evidence, frequently disjoint from the multimodal context and exhibiting significant hallucination.

In this work, we aim to address this problem by creating a scalable pipeline for generating high-quality, multi-hop, cross-modal reasoning data. Manually collecting such data is prohibitively expensive. While recent works have leveraged VLMs to scale up data collection [8, 15, 18, 56, 57], this approach has significant drawbacks for our target task.

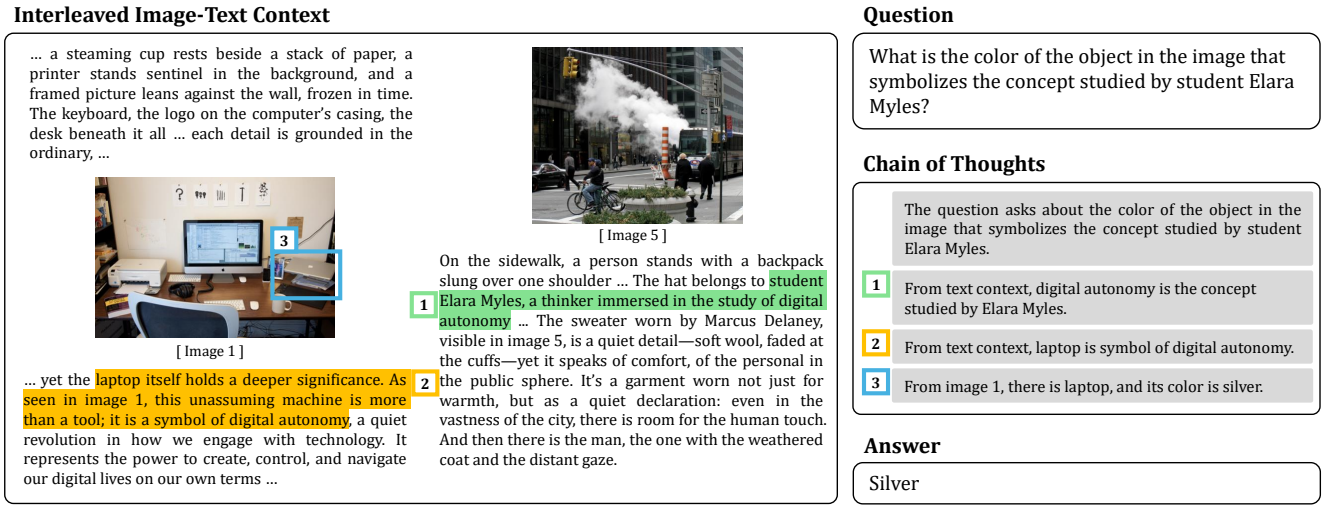


Figure 1. **Example from CRIT.** The CRIT dataset is designed to train and evaluate cross-modal multi-hop reasoning over interleaved image-text contexts. Each example combines multiple textual and visual sources that must be jointly interpreted to answer a compositional question. The colored highlights in the text indicate distinct evidence segments contributing to the reasoning chain, while the numbered boxes in the images and text correspond to the multi-step inference process outlined in the Chain of Thoughts panel. Together, they illustrate how CRIT requires connecting dispersed clues across modalities to reach a grounded answer.

Tasking a VLM to automatically generate complex reasoning questions is prone to the same grounding failures and hallucinations we seek to measure. Furthermore, this risks creating cyclical biases, where models are evaluated on data generated by the very same class of models.

To overcome these challenges, we propose a novel graph-based automatic data generation pipeline for interleaved image-text content. Our pipeline is built on several key properties: (i) First, we use graphs as a structured representation of content, capturing entities, attributes, and relationships that appear in either modality. These are derived from reliable sources, such as manually annotated captions; (ii) Second, this structured format allows us to programmatically sample sub-graphs, guaranteeing the presence of complex, multi-hop relationships between modalities; and (iii) Finally, given a sampled sub-graph, we use a model to generate a complex question that necessitates multi-hop reasoning to be solved. By design, our pipeline does not require a VLM at any stage of the question-generation process (LLM is sufficient), thus avoiding the aforementioned cyclical biases and grounding issues.

Using this pipeline, we construct a novel dataset called **CRIT** (**C**ross-modal multi-hop **R**easoning over interleaved **I**mage-**T**ext). Models trained on CRIT show improved cross-modal multi-hop reasoning, while our test set–built by our automatic pipeline and refined by human annotators–serves as a reliable benchmark for this capability.

To summarize, our key contributions are as follows:

1. We present an automatic graph-based data generation framework for cross-modal multi-hop reasoning across diverse domains, ranging from natural images to videos and text-rich sources such as scientific papers.
2. We use this pipeline to create the CRIT dataset where the test set has been verified manually by human raters. Along with the test set, we also benchmark a number of state-of-the-art VLMs and provide an in-depth analysis of common failure modes.
3. We further show that models trained on CRIT achieve significant gains in cross-modal multi-hop reasoning, with notable improvements on SPIQA and modest improvements on other standard multimodal benchmarks.

2. Related Work

Cross-Modal Multi-Hop Reasoning Benchmarking cross-modal multi-hop reasoning has received considerable attention in recent years [1, 4, 7, 12, 24, 45, 54, 62]. While these benchmarks aim to assess a model's ability to integrate information across modalities through multiple reasoning steps, many fall short of capturing the full complexity of the task. Datasets such as MMQA [48], for instance, are often solvable without visual inputs, indicating weak multimodal dependencies [25]. FCMR [25] introduces more challenging settings but still provides only QA pairs without explicit reasoning traces and remains limited in scope, focusing solely on the financial domain. Other datasets rely heavily on Wikipedia content, lack training splits, or depend on manual construction which limits scalability. Importantly, prior work seldom investigates cross-modal multi-hop reasoning over interleaved image–text inputs, where both modalities must be identified and com-

bined to locate relevant evidence and answer questions accurately. To address these limitations, we introduce CRIT, with a detailed comparison to existing cross-modal multi-hop datasets provided in Supp. Mat.

Interleaved Image-Text Datasets Interleaved image–text datasets [29, 33, 60, 64] are widely used for pretraining VLMs, but their weak cross-modal alignment limits their effectiveness for tasks requiring multi-hop reasoning across modalities. While they improve general multimodal capabilities, they do not guarantee strong performance when integrating complementary information from both modalities. Benchmarks such as SPIQA [42] and VEGA [63] attempt to address interleaved comprehension. However, SPIQA shows that models can perform well using only a single modality, suggesting limited cross-modal dependency. Similarly, VEGA builds on SciGraphQA [34], but much of the necessary information remains in the textual context, allowing many questions to be answered without visual input. Although these datasets remain useful, real-world scenarios often involve non-redundant, complementary information distributed across modalities. This highlights the need for datasets that better capture strong cross-modal interactions beyond loosely aligned interleaved corpora.

3. Data Generation for Cross-Modal Multi-Hop Reasoning

In this work we explore cross-modal multi-hop reasoning as shown in Figure 1. Obtaining suitable data for this task is non-trivial–even prior to annotation, the first challenge is identifying source data that contains high-quality image–text interleaved content with rich multimodal interactions. Once such content is collected, annotation requires identifying connected facts across modalities, formulating multi-hop questions and answers based on these facts, and ensuring that the supporting facts lead to a unique answer. One potential way to create such annotations is to prompt LLMs to automatically generate multi-hop questions and answers from raw image–text content. In practice, however, directly prompting the LLM to produce such examples without additional structure rarely yields valid cross-modal multi-hop data. The model often generates questions that rely on only one modality, collapse multi-hop reasoning into a shallow single-hop relation, or introduce hallucinated facts that are not grounded in the source content. We provide an illustrative failure case in Supp. Mat.

To address these challenges, we introduce a novel graph-based automatic data generation framework. Our framework operates in three stages, illustrated in Figure 2: **(1) constructing multimodal content graphs**, **(2) generating textual context** to accompany visual content, and **(3) generating question–answer pairs** that require cross-modal multi-hop reasoning based on the multimodal content graph. In the following subsections, we describe each stage in detail. We first describe how our framework works with images that already contain annotated scene graphs, and then describe extensions to our pipeline that enable us to use videos and scientific papers as well (Sec. 3.4).

3.1. Multimodal Content Graph Construction

The multimodal content graph lies at the core of our framework. It provides a structured representation of interleaved image–text content, consisting of entities, attributes, and relationships that appear in either modality. The graph encodes both intra-modal and cross-modal relationships.

Formulation of Multimodal Content Graphs. We model visual and textual content in interleaved image–text as a directed graph $G = (\mathcal{V}, \mathcal{E})$. Each node $v \in \mathcal{V}$ represents an entity, either a visual object in an image or a textual entity. An edge $e \in \mathcal{E}$ between nodes u and v is represented as a triple (u, v, r), where r denotes the relation between the two entities. The graph is further augmented with attributes associated with each entity when available.

Graph Construction through Content Augmentation. We begin with a dataset of images annotated with scene graphs, eliminating the need for image generation or potential errors in scene-graph recognition. Since our target is interleaved image–text content involving multiple images, we randomly sample one to six images from the annotated dataset (Figure 2 (a)). For each image, we use a rule-based method to retain only entities that can be uniquely identified by an attribute or by their relation to another entity, ensuring that each attribute or relation corresponds to exactly one entity (Figure 2 (b)). This avoids ambiguity during the QA generation phase, where multiple valid answers could otherwise arise. The scene graphs of the sampled images are then merged into a single content graph, which serves as the starting point for our framework. To augment this graph with textual information, we leverage the LLM to add new entities and relations. Specifically, for each image node in the graph, we prompt the LLM to generate a plausible entity and relation connected to that node (Figure 2 (c)). We then apply an additional round of prompting to create plausible relations among the newly added entities (Figure 2 (d)). These augmented entities and relations constitute the textual content; since they are not tied to a specific image, they serve as bridges that connect entities across different images through related textual nodes. The resulting graph provides a coherent multimodal context for evaluating cross-modal multi-hop reasoning, as it connects concepts drawn from multiple images and textual segments.

3.2. Textual Context Generation

After graph construction, each sample is represented by a unified graph structure, comprising both image and text nodes. For each image, we prompt the LLM with a subgraph to generate complementary textual context. We begin

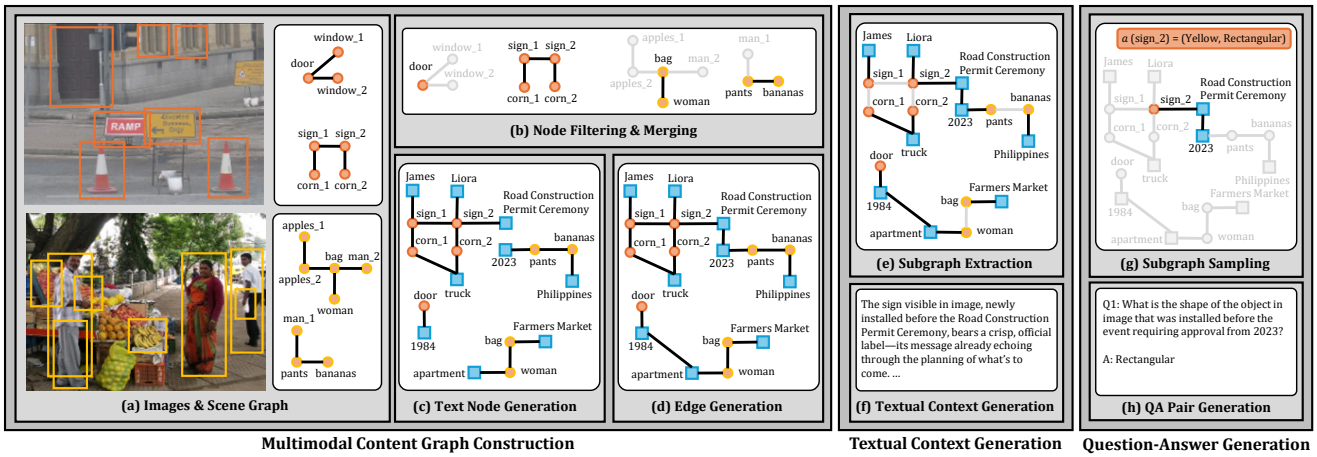


Figure 2. **Overall Process of Cross-Modal Multi-Hop QA Generation.** The procedure consists of three main stages. **(1) Multimodal Content Graph Construction:** images annotated with scene graphs are sampled (a), unique entities are filtered and merged (b), and text nodes connected with edges are generated via LLM prompting (c, d). **(2) Textual Context Generation:** subgraphs are extracted from the multimodal content graph (e) to produce complementary textual descriptions (f). **(3) Question–Answer Generation:** subgraphs are further sampled (g) to generate QA pairs requiring cross-modal multi-hop reasoning (h). Orange and yellow circular nodes represent visual entities originating from different images, while blue square nodes denote text nodes. Entities with identical names are distinguished by numerical subscripts (e.g., *apple_1*, *apple_2*).

by extracting the subgraph that includes the corresponding image nodes and all text nodes directly connected to them, along with the edges between those text nodes. When an edge connects two text nodes associated with different images, we randomly assign the edge to one image and expand its subgraph accordingly (Figure 2 (e)). Attributes of image nodes and inter-image relations are excluded, as these are intended to be inferred from the image during cross-modal reasoning. The generated passages therefore describe only the augmented text nodes, the relationships between them and their connections to the image nodes (Figure 2 (f)). Diverse narrative styles—such as stories, diary entries, or documentary-style scripts are used to provide linguistic variety. By combining the generated texts with the images, the framework yields complementary multimodal content that facilitates cross-modal multi-hop reasoning.

3.3. Question-Answer Pair Generation

Once the multimodal content is generated by augmenting images with textual content, we proceed to generate cross-modal multi-hop QA pairs. To this end, we sample a subgraph from the multimodal content graph that represents a chain of facts within the content (Figure 2 (g)). Since our goal is cross-modal reasoning, we ensure that the sampled subgraph includes nodes from both images and text. To guarantee multi-hop reasoning, we further restrict the subgraphs to have $1 \leq h \leq 5$ edges, such that answering the resulting questions requires reasoning over multiple connected facts. Finally, we require the terminal node in each chain to come from an image. For $h = 1$, we sample such subgraphs only when the terminal image node possesses an attribute and use this attribute as the answer to ensure the reasoning remains two-hop. For $h > 1$, we select either its entity name or one of its attributes as the answer, grounding the final reasoning step in the image content. This strategy improves the reliability of evaluation by preventing the model from guessing answers based solely on textual biases without performing the correct cross-modal reasoning.

Given a sampled subgraph (serialized as JSON) and its designated answer, we prompt the LLM to generate a question under the following constraints: (i) the answer to the generated question must match the provided target answer; (ii) solving the question must require reasoning over the entire chain of facts in the subgraph, thereby enforcing cross-modal multi-hop reasoning; and (iii) intermediate entities should not be mentioned explicitly in the question but instead be recoverable only through multi-hop inference (Figure 2 (h)). While the QA pair provides the basis for evaluating cross-modal multi-hop reasoning, we additionally generate a CoT trace [55] to be used for training and evaluation. Since the subgraph corresponding to each QA already encodes the complete reasoning chain, we can leverage it to prompt the LLM to produce a CoT response for the generated question. This offers extra supervision by explicitly indicating where each piece of information should be retrieved from within the multimodal context.

Once the QA samples with CoT are generated, we apply the LLM-based staged filtering process to ensure quality. First, we discard questions that explicitly mention intermediate entities from the sampled subgraph, thereby ensur-

ing multi-hop reasoning by eliminating potential shortcuts in the reasoning chain. Next, we filter out questions that can be answered using a single modality. To perform this test, we provide LLMs with the nodes and edges corresponding to each modality; since visual content is translated into structured text through the graph representation, a text-only LLM is sufficient. We use three different LLMs and remove a question if all three predict the correct answer using only one modality (either text or visual). Finally, we prune CoT responses that are excessively verbose by limiting the output to a maximum of ten sentences.

3.4. Extensions to Videos and Text-Rich Sources

The data creation pipeline described above is limited in two aspects: (i) first, it relies on images that have graph annotations, and (ii) the images are independent and unrelated to one another, as well as being limited to the natural image domain. We therefore aim to extend the pipeline to settings lacking pre-existing graph information by first deriving graph representations from alternative annotations and then applying the same pipeline. We also seek to incorporate text-rich visuals such as tables, diagrams, and charts that frequently occur alongside textual context in real-world documents [16, 22, 61]. To address these gaps, we extend our pipeline to two additional data sources: (i) still frames from videos, where multiple coherent frames naturally share contextual information, and (ii) scientific papers, which provide abundant text-rich images.

For video frames, the main challenge is the absence of scene-graph annotations. To address this, we leverage dense video captioning datasets, where long-form videos are annotated with temporally localized event descriptions. We select the frames with the highest CLIP similarity [43] to each caption to ensure that the visual content aligns closely with the description. Although these captions may be short, they provide sufficient content information to be converted into partial scene graphs—similar to the manual scene graph annotation process in [27], where captions are collected and then transformed into graphs. Concretely, we prompt the LLM to convert these descriptions into graphs. Sampled frames from a single video may share entities and relations, and therefore we incorporate this into both the prompting design and the resulting graphs for coreference resolution. After constructing a scene graph across coherent frames, the rest of the pipeline proceeds as before, except that the textual context is now generated once for the entire image set rather than separately for each image.

In the case of text-rich images, scientific papers inherently provide high-quality interleaved image–text content. Because well-aligned textual descriptions already exist, generating complementary textual context is unnecessary. Nonetheless, constructing QA pairs would benefit from a graph-based representation. To this end, we convert the multimodal content—paragraphs, figures, and tables—into a unified graph structure, treating both figures and rendered tables as images. Given a paper containing textual content with associated figures and tables, we first construct a content graph from paragraphs that do not reference any visual elements using the LLM. For paragraphs that reference figures or tables, we prompt the LLM to identify visual entities and relations. Numerical entities, comparisons, and explicit visual descriptions are elevated to visual nodes in the graph. To enforce cross-modal reasoning, the corresponding sentences are then removed from the text body once they are marked as visual. Finally, once the corresponding multimodal content graph is constructed, QA generation is performed following the same base pipeline. In this case, we restrict the subgraph to have $1 \leq h \leq 4$ edges, as their relations are more complex and larger chains often lead the LLM to generate overly complicated or unnatural questions. Further implementation details and prompt templates can be found in Supp. Mat.

4. CRIT

In this section, we describe the implementation details used to create CRIT (sec. 4.1), and dataset statistics (sec. 4.2).

4.1. Data Sources and Models

Data Sources CRIT is based on three primary data sources, corresponding to three distinct domains: natural image, video frame, and scientific paper. Specifically, the natural image domain is constructed from GQA [19], the video frame domain from ActivityNet Captions [26], and the scientific paper domain from SPIQA [42]. We follow the official train/test splits of each source to create the CRIT training set and benchmark. From GQA, we utilize images and their associated scene graphs, sampled and processed following the pipeline described in Section 3.1 to Section 3.3. Beyond these, CRIT additionally incorporates video frames and captions from ActivityNet Captions, and also includes interleaved image–text pairs extracted from scientific papers in SPIQA, as described in Section 3.4. Importantly, no QA pairs from these datasets are directly used in CRIT; all QA annotations are newly generated through our pipeline. Further implementation details and all prompts used are included in Supp. Mat.

Models Qwen3-30B-A3B-Instruct-2507 [51] was used at every stage of the data generation pipeline. For the final filtering of questions and answers, we used Qwen3-30B-A3B-Instruct-2507 [51], Gemma-3-27b-it [49], and Mistral-Small-3.2-24B-Instruct [3].

4.2. Dataset Analysis

CRIT is divided into train and test sets derived from the official splits of each data source. For the SPIQA source, due

Table 1. **CRIT Dataset Statistics.** CRIT comprises three domains: natural image (NI), video frame (VF), and scientific paper (SP). This table reports the key statistics for each domain.

Split	Statistic	NI	VF	SP
Train	# Samples	49,159	8,022	27,018
	Avg. Images/Sample	3.8	3.5	3.7
	Avg. Text Tokens/Sample	2,442	649	5,443
	# QA	153,781	16,071	99,615
	# 2-hop	109,735	8,061	58,429
	# 3-hop	12,271	6,042	28,385
	# 4-hop	12,592	849	12,801
	# 5-hop	19,183	1,119	–
Test	# Samples	792	332	322
	Avg. Images/Sample	4.4	5.2	7.7
	Avg. Text Tokens/Sample	2,930	802	5,338
	# QA	858	366	383
	# 2-hop	597	134	245
	# 3-hop	71	163	110
	# 4-hop	74	34	28
	# 5-hop	116	35	–

to the limited size of its official test set, we sample a subset from its training split to form the CRIT test split for the scientific paper domain, ensuring no overlap between train and test samples. For the test split, we recruited 12 English-proficient annotators majoring in STEM disciplines to manually verify and filter samples that did not adhere to the guidelines. In contrast, the train set is automatically constructed without human curation, and may therefore contain some degree of noise. Each CRIT data sample comprises a set of one or more images, an accompanying textual context, one or multiple QA pairs, and corresponding CoT reasoning traces. During training, samples with multiple QA pairs are concatenated to form multi-turn conversations while test samples are restricted to single-turn QA pairs. In total, CRIT contains 84,199 train samples and 1,446 test samples, corresponding to 269,467 QA pairs in the train split and 1607 QA pairs in the test split. Table 1 summarizes the detailed statistics across data sources.

5. Experiments

5.1. Experimental Setup

Base Models We evaluate both open-source and proprietary models. For open-source models, we evaluate Qwen2.5-VL [6], LLaVA-Onevision [32], Idefics2 [30], Intern2.5-VL [10], Phi3.5-Vision [2]. For proprietary models we use GPT-4o, GPT-4o-mini [20] and Gemini-2.0-flash [50].

Training Settings We perform supervised fine-tuning using LoRA [17] on VLMs that support multi-image input, specifically Qwen2.5-VL-7B and Idefics2-8B on the CRIT training set. Each training sample is included in two formats: one containing only the final answer and another including the full CoT reasoning trace. We use both formats to enable the models to generate concise answers as well as step-by-step reasoning responses.

Table 2. **Results on CRIT.** Performance comparison of proprietary, open-source, and fine-tuned VLMs across natural image (NI), video frame (VF), and scientific paper (SP). Models fine-tuned on CRIT are highlighted with a gray background. Open-source models are grouped by model size for clearer comparison. *# Prm* denotes the number of model parameters.

		NI		VF		SP	
Model	**# Prm**	EM	F1	EM	F1	EM	F1
Proprietary Models							
Gemini 2.0 Flash [50]	-	30.4	32.7	30.6	35.3	11.5	13.7
GPT-4o [20]	-	35.1	37.7	32.0	38.9	8.4	14.0
GPT-4o-mini [20]	-	24.7	25.5	28.7	33.3	9.7	13.1
Open-Source Models							
Phi3.5-Vision [2]	4B	21.5	23.0	24.0	26.5	3.1	5.9
LLaVA-Onevision [32]	7B	31.0	32.0	33.3	37.4	7.1	10.1
Qwen2.5-VL [6]	7B	28.3	29.1	24.0	27.8	6.8	9.6
Qwen2.5-VL$_{\text{CRIT}}$	7B	**58.6**	**59.5**	**38.8**	**42.2**	**15.9**	**22.5**
Intern2.5-VL [10]	8B	26.7	27.9	30.1	34.8	5.5	9.8
Idefics2 [30]	8B	18.3	18.9	20.2	25.2	3.4	6.1
Idefics2$_{\text{CRIT}}$	8B	54.1	54.9	31.2	33.9	12.3	20.2
Qwen2.5-VL [6]	72B	38.0	39.4	30.1	33.9	9.4	12.3

Evaluation Details Evaluation is conducted on the CRIT benchmark, which covers three domains: natural images, videos, and scientific papers. Two evaluation settings are considered: (1) *Direct Answer*, where the model is prompted to produce an answer directly without reasoning steps, and (2) *CoT*, where the model is instructed to reason step-by-step before providing the final answer. Model performance is measured using Exact Match (EM) and F1 scores on the final answer [44]. We observe similar trends across both evaluation settings; therefore, unless otherwise specified, results are reported using CoT prompting by default, with direct answer results provided in Supp. Mat.

5.2. Results on CRIT

Table 2 presents the performance of various VLMs on the CRIT benchmark under the CoT evaluation setting. Overall, all models perform poorly, including proprietary systems. Although proprietary or larger open-source models such as GPT-4o and Qwen2.5-VL-72B achieve moderately better results than smaller models, their absolute performance remains low. This reveals that, despite recent advances in multimodal reasoning, existing models still have limited capability in cross-modal multi-hop reasoning, highlighting the substantial challenges posed by CRIT. When fine-tuned on the CRIT training set, both Qwen2.5-VL$_{\text{CRIT}}$ and Idefics2$_{\text{CRIT}}$ demonstrate substantial improvements over their corresponding baselines across all domains, surpassing even proprietary models in both EM and F1. Nonetheless, the scientific-paper domains remain

Table 3. **Performance on Cross-Modal Reasoning Benchmarks.** Performance comparison of zero-shot Idefics2-8B and the same model fine-tuned on Mantis-Instruct alone or jointly with CRIT, where Mantis-Instruct and CRIT have a 5:1 dataset size ratio. *# Samples* indicates the total number of training samples in the corresponding fine-tuned dataset.

		SPIQA				VEGA		MMQA		FCMR
Fine-tuned Dataset	**# Samples**	METEOR	ROUGE-L	CIDEr	BERTScore F1	ROUGE-L	BLEU	EM	F1	F1
No Fine-tuning	-	1.13	1.10	1.03	28.23	31.4	6.7	27.4	31.7	40.5
Mantis-Instruct	837k	3.60	10.17	23.83	42.83	29.5	5.1	27.3	30.9	44.9
Mantis-Instruct+CRIT	1005k	**10.53**	**22.77**	**67.93**	**55.80**	**35.1**	**7.1**	**30.0**	**33.8**	**50.5**

Table 4. **Performance on Standard Multimodal Benchmarks.** Performance comparison of zero-shot Idefics2-8B and the same model fine-tuned on Mantis-Instruct alone or jointly with CRIT, where Mantis-Instruct and CRIT have a 5:1 dataset size ratio. *# Samples* indicates the total number of training samples in the corresponding fine-tuned dataset.

Fine-tuned Dataset	**# Samples**	**MuirBench**	**BLINK**	**MP-DocVQA**	**MMStar**	**MME**	**SeedBench**	**ChartQA**
No Fine-tuning	-	26.2	45.2	46.7	30.2	1792	26.4	26.4
Mantis-Instruct	837k	**33.1**	45.7	48.2	47.2	1909	**67.3**	63.8
Mantis-Instruct+CRIT	1005k	32.6	**47.6**	**49.8**	**48.3**	**1910**	**67.3**	**64.9**

particularly challenging. For scientific papers, long and complex textual contexts hinder precise alignment between visual and textual information. As a result, even trained models continue to struggle with maintaining coherent reasoning in these settings. Model performance across different hop counts is included in Supp. Mat.

5.3. Results on Other Benchmarks

We further investigate the impact of fine-tuning on CRIT across several benchmarks that evaluate the general multimodal capabilities of VLMs, including SPIQA [42], VEGA [63], MMQA [48], FCMR [25], MuirBench [53], BLINK [14], MP-DocVQA [52], MMStar [9], MME [13], SeedBench [31] and ChartQA [38]. The first four benchmarks primarily focus on complex cross-modal reasoning, while the remaining benchmarks represent standard multimodal evaluation suites covering general capabilities across both single-image and multi-image settings. All evaluations follow the official settings and metrics defined for each benchmark. Specifically, we compare the pretrained Idefics2 model and its variants fine-tuned with and without CRIT, in addition to Mantis-Instruct [23], across these benchmarks. Mantis-Instruct is a widely adopted public dataset known for improving general instruction-following capabilities over multiple images in VLMs. We jointly fine-tune on Mantis-Instruct and CRIT to preserve the model's existing instruction-following and multimodal capabilities while allowing CRIT to act as a complementary dataset that provides focused supervision for cross-modal reasoning.

Results on cross-modal reasoning benchmarks are reported in Table 3. These benchmarks are derived from diverse human-authored sources, such as academic publications, Wikipedia articles, and financial reports, reflecting a wide range of real-world reasoning scenarios. The pretrained model without fine-tuning performs poorly, reflecting limited exposure to cross-modal reasoning during pretraining. Fine-tuning on Mantis-Instruct alone improves performance, consistent with prior work [23], while incorporating CRIT yields substantial gains across all benchmarks. On SPIQA, fine-tuning with CRIT more than doubles performance across METEOR, ROUGE-L, and CIDEr. Beyond SPIQA, training with CRIT yields consistent improvements across other benchmarks, including relative gains of 19%/39% (ROUGE-L/BLEU) on VEGA, 10% (EM/F1) on MMQA, and 12% (F1) on FCMR, demonstrating substantial improvements in general cross-modal reasoning. Note that although scientific paper data from SPIQA are used as contextual examples during dataset construction, no SPIQA question–answer pairs are included in either data construction or fine-tuning, and the SPIQA validation and test sets are never used at any stage.

Table 4 presents results on broader multimodal benchmarks. Fine-tuning with CRIT preserves performance on MME and SeedBench, while improving over the baseline on BLINK, MP-DocVQA, MMStar, and ChartQA. Performance on MuirBench remains comparable to the model trained with Mantis-Instruct alone. Overall, CRIT consistently enhances performance across both specialized cross-modal reasoning benchmarks and general multimodal evaluations. These findings suggest that training on more challenging cross-modal multi-hop reasoning examples improves the generalization and transferability of VLMs without degrading their existing capabilities.

5.4. Extending the CRIT Dataset

We study whether scaling CRIT improves performance by augmenting its training data, based on fine-tuning Qwen2.5-VL-7B. The original CRIT dataset relies on human-labeled

Table 5. **Results on CRIT Extension.** Performance comparison of zero-shot Qwen2.5-VL-7B and the same model fine-tuned on the original CRIT training set and an extended version augmented with additional samples. *# Samples* indicates the total number of training samples in the corresponding fine-tuned dataset.

		NI		VF		SP	
Fine-tuned Dataset	**# Samples**	EM	F1	EM	F1	EM	F1
No Fine-tuning	–	28.3	29.1	24.0	27.8	6.8	9.6
CRIT	84k	58.6	59.5	38.8	42.2	15.9	22.5
CRIT Augmented	210K	**62.6**	**63.7**	**45.6**	**49.1**	**16.7**	**22.8**

data sources in the natural image and video frame domains, where scene graphs from GQA and captions from ActivityNet Captions are manually annotated sources. As a more scalable alternative, we extend CRIT by constructing additional training samples using annotations produced by off-the-shelf models, which are inherently noisier than human-curated ones. This allows us to examine whether such noisy data can complement manually annotated samples and enhance cross-modal reasoning performance.

To this end, we augment the CRIT training set in both natural image and video domains. For natural images, we use images from Open Images [28] and generate scene graphs using EGTR [21], supplemented by object attributes generated using Qwen2.5-VL-7B. For videos, we use video frames extracted from Charades [47], DiDeMo [5], and VidOR [46], and generate video captions using Qwen2.5-VL-32B. Importantly, the same CRIT data generation pipeline is applied to these automatically annotated sources, ensuring consistency with the original data generation pipeline. This augmentation produces approximately 100K additional samples for natural images and 25K for videos.

As shown in Table 5, augmentation consistently improves performance across all domains. The largest improvement appears in the video domain, likely due to its smaller original dataset, which benefits more from additional samples. The performance of the scientific paper domain also improves even though augmentation is applied only to the natural image and video domains, suggesting cross-domain transfer of cross-modal reasoning capabilities. Overall, these results demonstrate that model-generated annotations provide a scalable and effective way to enhance cross-modal reasoning.

6. Error Analysis

We conducted an error analysis to identify the key bottlenecks of current VLMs in cross-modal multi-hop reasoning tasks. Specifically, we use 75 randomly selected incorrect predictions from GPT-4o, each of which is manually verified to ensure the error is genuine before analysis. We categorize these failures into four types: 1) **Evidence Localization Error**: The model fails to correctly locate or ground the relevant evidence needed to answer the question. 2) **Visual Perception Error**: The model correctly identifies the relevant image but hallucinates or misperceives its visual content. 3) **Textual Comprehension Error**: The model hallucinates or misinterprets information within the textual context. 4) **Answer Generation Error**: The model correctly gathers relevant information but produces an incorrect final answer due to flawed reasoning. Figure 3 shows the distribution of these errors. Among these, evidence localization error is the most prevalent, accounting for 55% of all observed failures. These cases involve the model selecting the wrong image or text paragraph, or seeking information in the wrong modality. This highlights that our benchmark effectively reveals the limitations of current VLMs in cross-modal multi-hop reasoning, where models must integrate and reason over information from both visual and textual sources. The second most frequent error type is visual perception error, which occurs roughly four times more often than textual comprehension error, suggesting that models remain comparatively robust in language understanding. The least common category is answer generation error, implying that when all relevant evidence is correctly identified, models are generally capable of producing the correct final answer. Qualitative examples of each error type are provided in Supp. Mat.

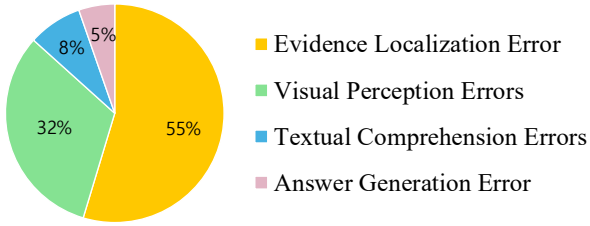


Figure 3. **Error categories and their distribution across 75 randomly sampled GPT-4o responses on CRIT.**

7. Conclusion

We explored the challenge of cross-modal multi-hop reasoning, where solving a task requires weaving together complementary evidence from interleaved text and images. To advance this capability, we introduced CRIT, built through a novel graph-based automatic pipeline that yields scalable, high-quality training and evaluation data. Fine-tuning on CRIT delivers notable gains across diverse benchmarks without degrading their existing capabilities, and our augmentation experiments demonstrate a practical path to further scaling. We believe CRIT provides a foundation for developing models with more robust and grounded multimodal reasoning.

Acknowledgements

This work was supported by the IITP grants (RS-2025-02653113, IITP-2026-RS-2020-II201819, IITP-2026-RS-2024-00436857, RS-2024-00398115, IITP-2026-RS-2025-02304828), the NRF grant (RS-2025-23523979) and the KOCCA grant (RS-2024-00345025) funded by the Korean government (MSIT, MSCT).

References

[1] Amirhossein Abaskohi, Spandana Gella, Giuseppe Carenini, and Issam H Laradji. Fm2ds: Few-shot multimodal multihop data synthesis with knowledge distillation for question answering. *arXiv preprint arXiv:2412.07030*, 2024. 2

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 6

[3] Mistral AI. Mistral small 3.1, 2025. Accessed: 2025-09-25. 5

[4] Mubashara Akhtar, Chenxi Pang, Andreea Marzoca, Yasemin Altun, and Julian Martin Eisenschlos. Tanq: An open domain dataset of table answered questions. *Transactions of the Association for Computational Linguistics*, 13: 461–480, 2025. 2

[5] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 8

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[7] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022. 1, 2

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 1

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 1, 7

[10] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6

[11] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024. 1

[12] Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. Wikimixqa: A multimodal benchmark for question answering over tables and charts. *arXiv preprint arXiv:2506.15594*, 2025. 2

[13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 7

[14] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *ArXiv*, abs/2404.12390, 2024. 1, 7

[15] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024. 1

[16] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 5

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[18] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[21] Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. Egtr: Extracting graph from transformer for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24229–24238, 2024. 8

[22] Mengzhao Jia, Wenhao Yu, Kaixin Ma, Tianqing Fang, Zhihan Zhang, Siru Ouyang, Hongming Zhang, Dong Yu, and Meng Jiang. Leopard: A vision language model for text-rich multi-image tasks. *arXiv preprint arXiv:2410.01744*, 2024. 5

[23] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 7

[24] Jihyung Kil, Farideh Tavazoee, Dongyeop Kang, and Joo-Kyung Kim. Ii-mmr: Identifying and improving multi-modal multi-hop reasoning in visual question answering. *arXiv preprint arXiv:2402.11058*, 2024. 2

[25] Seunghee Kim, Changhyeon Kim, and Taeuk Kim. Fcmr: Robust evaluation of financial cross-modal multi-hop reasoning. *arXiv preprint arXiv:2412.12567*, 2024. 2, 7

[26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In

Proceedings of the IEEE international conference on computer vision, pages 706–715, 2017. 5

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5

[28] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 8

[29] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 3

[30] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024. 6

[31] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 7

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[33] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024. 3

[34] Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*, 2023. 3

[35] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1

[36] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1

[37] Haohao Luo, Ying Shen, and Yang Deng. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9355–9367, Singapore, 2023. Association for Computational Linguistics. 1

[38] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1, 7

[39] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *ArXiv*, abs/2408.02718, 2024. 1

[40] Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, et al. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*, 2024. 1

[41] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, et al. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*, 2025. 1

[42] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833, 2024. 3, 5, 7

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[44] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 6

[45] Revant Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, et al. Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11200–11208, 2022. 2

[46] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019. 8

[47] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European conference on computer vision*, pages 510–526. Springer, 2016. 8

[48] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021. 2, 7

[49] Gemma Team. Gemma 3. 2025. 5

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, and et al. Katie Millican. Gemini: A family of highly capable multimodal models, 2025. 6

[51] Qwen Team. Qwen3 technical report, 2025. 5

[52] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 7

[53] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. *ArXiv*, abs/2406.09411, 2024. 1, 7

[54] Haoran Wang, Aman Rangapur, Xiongxiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. Piecing it all together: Verifying multi-hop multimodal claims. *arXiv preprint arXiv:2411.09547*, 2024. 2

[55] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 4

[56] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 1

[57] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 1

[58] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 1

[59] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 1

[60] Wenqi Zhang, Hang Zhang, Xin Li, Jiashuo Sun, Yongliang Shen, Weiming Lu, Deli Zhao, Yueting Zhuang, and Lidong Bing. 2.5 years in class: A multimodal textbook for vision-language pretraining. *arXiv preprint arXiv:2501.00958*, 2025. 3

[61] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 5

[62] Bowen Zhao, Tianhao Cheng, Yuejie Zhang, Ying Cheng, Rui Feng, and Xiaobo Zhang. Ct2c-qa: Multimodal question answering over chinese text, table and chart. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3897–3906, 2024. 2

[63] Chenyu Zhou, Mengdan Zhang, Peixian Chen, Chaoyou Fu, Yunhang Shen, Xiawu Zheng, Xing Sun, and Rongrong Ji. Vega: Learning interleaved image-text comprehension in vision-language large models. *arXiv preprint arXiv:2406.10228*, 2024. 3, 7

[64] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023. 3