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

Supplementary Material

A. Limitations

Despite the strong performance gains achieved by models fine-tuned with CRIT—constructed via a graph-based automatic data generation pipeline—several limitations persist. First, the dataset is entirely synthesized by the LLM. Although the graph-based formulation imposes structural constraints and leverages human-annotated sources as initialization sources, the resulting QA pairs and reasoning chains may still reflect biases or stylistic patterns inherent to the generator model. In addition, the LLM-generated textual contexts may contain fictional or non-grounded content. That said, when evaluated on existing cross-modal reasoning benchmarks derived from diverse human-authored sources (e.g., SPIQA, VEGA, MMQA, FCMR), models trained with CRIT consistently demonstrate performance gains, suggesting that the learned capabilities generalize beyond the synthetic data distribution. Nevertheless, such evaluations may not fully surface subtle generator-specific biases, which could still be present but remain unobserved under current benchmarks. Second, the multi-stage nature of the data generation pipeline introduces potential error accumulation, as LLMs are employed throughout the process. Imperfections at earlier stages may propagate and amplify in subsequent steps. Finally, the pipeline incurs non-trivial computational overhead, which may pose challenges for efficiency and resource usage.

B. Comparison with Prior Work

Table 2 summarizes the key differences between CRIT and existing datasets based on cross-modal reasoning. Although datasets such as SPIQA [19] and VEGA [30] adopt interleaved image-text inputs, and others like MMQA [23] and WebQA [4] can be reformulated into this format, the majority of their questions are still solvable using a single modality. As shown in Table 1, this limits their ability to probe genuine cross-modal reasoning. CRIT is explicitly designed to overcome this limitation. Every question requires multi-hop, cross-modal inference, compelling models to integrate complementary visual and textual evidence rather than rely on isolated cues.

A second differentiator is domain diversity. Existing datasets commonly rely on narrow sources, such as Wikipedia [4, 11, 23] or specialized scientific and financial corpora [10, 14, 19, 30]. In contrast, CRIT draws from a broader and more heterogeneous set of public data, enabling evaluation of general-purpose multimodal understanding. Finally, CRIT emphasizes scalability and interpretability. Using the LLM-driven pipeline with manual verification of the test split, we construct 271k high-quality QA pairs. Our graph-based generation process additionally provides the exact subgraph used to form each question, yielding transparent reasoning traces that support training, diagnostics, and deeper analysis of model behavior.

Table 1. **Ablation on Image Input.** We evaluate GPT-4o on benchmarks formulated as interleaved image–text inputs with and without images. Performance remains strong on existing benchmarks but drops sharply on CRIT, indicating that CRIT is designed to require cross-modal reasoning. M: METEOR, R-L: ROUGE-L, C: CIDEr, B-F1: BERTScore F1, B: BLEU.

	SPIQA				VEGA		MMQA		CRIT	
Image	M	R-L	C	B-F1	R-L	B	EM	F1	EM	F1
✓	17.3	30.8	88.0	57.8	40.1	12.3	48.8	56.0	28.0	32.3
✗	11.8	19	36.6	49.9	40.3	12.7	42.9	49.6	10.5	12.5

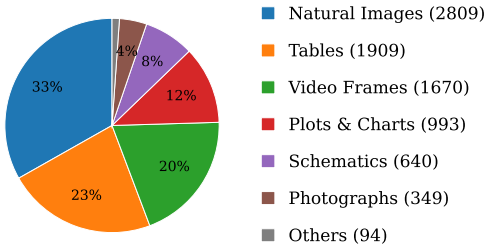


Figure 1. **Image Distribution**

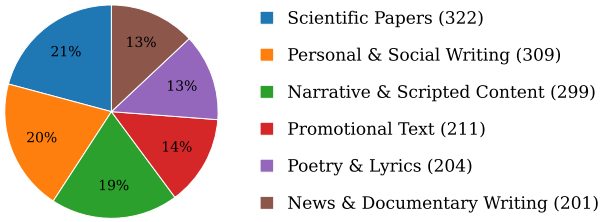


Figure 2. **Text Distribution**

C. Domain Distribution of CRIT

Figures 1 and 2 illustrate the distribution of image and text data across domains in the CRIT benchmark, demonstrating broad coverage with samples spanning diverse image cate-

Table 2. **Comparison of our dataset with prior cross-modal reasoning works.** CRIT targets cross-modal multi-hop reasoning over interleaved image-text inputs, constructed through a graph-based automatic pipeline spanning diverse data sources, and provides 271k QA pairs together with explicit reasoning traces.

Dataset	Interleaved Image-Text	Multi Hop	Diverse Data Source	Data Generation	Reasoning Trace	Num QA
ManyModalQA [11]	✗	✗	✗	Human experts	✗	10.1k
CT2C-QA [29]	✗	✗	✗	LLMs + Human experts	✗	10k
MME-Finance [10]	✗	✗	✗	LLMs + Human experts	✗	2.3k
WebQA [4]	✗	✓	✗	Human experts	✗	46.7k
MuMuQA [20]	✗	✓	✗	Human experts	✗	1.4k
MMQA [23]	✗	✓	✗	Template + Human experts	✗	30k
FCMR [14]	✗	✓	✗	LLMs + Human experts	✗	2.2k
SPIQA [19]	✓	✗	✗	LLMs + Human experts	✓	270k
VEGA [30]	✓	✗	✗	LLMs + Human experts	✗	286k
CRIT	✓	✓	✓	LLMs + Human experts	✓	271k

gories and text genres.

D. Data Generation Detail for Video and Text-Rich Sources

This section provides further details on (1) how video captions are used to form a unified graph across sampled frames, (2) how text-rich sources are converted into multimodal graphs, and (3) how textual context is refined when paragraphs refer to visual elements.

For video sources, the full set of captions along with their temporal ranges is provided to the LLM. The model identifies all entities mentioned throughout the video, extracts their attributes, and performs coreference resolution to merge entities appearing across multiple frames. Frame-level relations are then inferred, capturing interactions and events described in each segment. Compared with natural images, the resulting video graphs are more action-centric and reflect the temporal structure of the content. The prompt used for caption-to-graph conversion is shown in Figure 19.

Scientific papers are processed directly from their raw TeX sources. After segmenting the text into paragraphs, we prompt the LLM to extract entities and textual relations for paragraphs that do not reference any figures or tables, using the prompt shown in Figure 20. For paragraphs that do contain such references, we extract only relations grounded in the referenced visual content—e.g., numerical results ("Model A achieves 92% accuracy"), comparative statements ("Method B outperforms Method C"), or visual attributes ("The blue curve increases sharply"), following the prompt in Figure 21.

To ensure that questions require genuine cross-modal reasoning, paragraphs referencing figures or tables undergo sentence-level filtering. After visual relations are extracted, we embed each sentence in the paragraph using SBERT (all-MiniLM-L6-v2) [21] and compute cosine similarity with the extracted visual relations. Sentences that closely overlap with the visual content are removed according to a fixed threshold, while all remaining sentences are retained and used as the textual context. Paragraphs without visual references are left unchanged. This yields textual context that complements rather than duplicates the visual evidence, promoting reasoning chains that span modalities.

E. Failure Cases of Data Generation Without Graphs

Figures 3 and 4 illustrate failure cases that arise when prompting the LLM to directly produce cross-modal multi-hop QA pairs from raw image–text inputs without any additional structural guidance. For each example, one image and its associated paragraph are randomly sampled from the interleaved image-text context, and the LLM is asked to produce a QA pair requiring cross-modal multi-hop reasoning given the caption of the sampled image and the paragraph. Figure 3 shows a case where the generated question appears cross-modal but is actually solvable using a single modality. The model implicitly incorporates visual information into the question itself, making the textual paragraph alone sufficient for answering. Figure 4 shows a question that is syntactically framed as multi-hop, yet relies on only a single reasoning step. These cases underscore typical shortcomings observed when prompting LLMs without additional structural constraints, motivating the graph-based approach introduced in the main paper.

F. Ablation Study on Data Generation Pipeline

We conduct a retrospective ablation study to assess the contribution of key components in our data generation pipeline. Experiments are performed on 100 randomly sampled instances from the CRIT test set, using GPT-5 [22] as an automatic evaluator. We first evaluate the role of cross-image edges in the multimodal content graph. These edges connect entities across different images via textual nodes and enable multi-hop reasoning that spans multiple images. When removing all cross-image edges, we find that 33% of the questions become unanswerable. This demonstrates that

Table 3. **Ablation of Fine-Tuning Data Composition** Performance comparison of zero-shot Idefics2-8B and the same model fine-tuned on different datasets. We compare fine-tuning on Mantis-Instruct alone and jointly with CRIT or SPIQA. Mantis-Instruct and CRIT follow a 5:1 sampling ratio, and all joint settings contain the same total number of training samples. Joint training with CRIT provides the strongest overall improvements, while size-matched SPIQA yields notably smaller gains, underscoring the contribution of CRIT. M: METEOR, R-L: ROUGE-L, C: CIDEr, B-F1: BERTScore F1, B: BLEU, Acc: Accuracy.

		SPIQA				VEGA		MMQA		FCMR	MuirBench	BLINK	MP-DocVQA
Fine-tuned Dataset	**# Samples**	M	R-L	C	B-F1	R-L	B	EM	F1	F1	Acc	Acc	Acc
No Fine-tuning	-	1.13	1.10	1.03	28.23	31.4	6.7	27.4	31.7	40.5	26.2	45.2	46.7
Mantis-Instruct	837k	3.60	10.17	23.83	42.83	29.5	5.1	27.3	30.9	44.9	33.1	45.7	48.2
Mantis-Instruct+CRIT	1005k	**10.53**	**22.77**	67.93	**55.80**	**35.1**	**7.1**	**30.0**	**33.8**	**50.5**	32.6	**47.6**	**49.8**
Mantis-Instruct+SPIQA	1005k	10.0	22.47	**71.7**	54.57	28.4	4.0	29.2	32.3	47.5	**34.9**	46.4	48.8

such connections are essential for maintaining valid cross-modal reasoning paths. Next, we examine the impact of graph-based QA generation. We compare QA pairs generated using our graph-structured pipeline with those produced via direct prompting without explicit graph guidance. Using GPT-5 as a judge with the prompt shown in Figure 11, graph-based QA pairs are preferred in 77% of cases. This preference indicates that graph-based generation produces questions that are more coherent, better grounded in the underlying multimodal content, and more consistent with the intended multi-hop reasoning chains. Overall, the results highlight the importance of both cross-image connectivity and graph-based generation in producing high-quality cross-modal multi-hop reasoning data.

G. Benchmarks and Evaluation

This section provides additional details on the benchmarks and clarifies how evaluation scores are reported.

SPIQA SPIQA [19] is an open-ended QA dataset tailored for interpreting complex figures and tables in scientific research articles across diverse subfields of computer science. Each instance contains interleaved image–text inputs, where figures and tables are provided as images accompanied by their captions. The test set comprises three splits. We report the average performance across them.

VEGA VEGA [30] is a dataset designed for interleaved image-text comprehension, which requires models to identify relevant visual and textual regions within a complex multimodal context and generate accurate answers to the given questions. The test set includes two splits differing in input token length. We report averaged results over both.

MMQA MMQA [23] is a multi-source question-answering benchmark constructed from Wikipedia, requiring reasoning over text passages, tables, and images. The dataset includes both single-hop and multi-hop questions, with some cases demanding cross-modal integration to infer the correct answer. In our setting, we provide all text, image, and table inputs at once in an interleaved format. We report multi-hop performance on the validation set.

FCMR FCMR [14] evaluates multimodal reasoning in the financial domain, combining textual reports, tables, and charts. Each instance provides three candidate statements whose verification requires evidence aggregation across modalities. The official all-or-nothing scoring scheme counts a prediction as correct only when the exact set of true statements is returned; in practice, this criterion is overly strict and leads models to behave nearly randomly. To provide a more informative assessment, we instead report per-statement F1 scores computed using the ground-truth labels, treating each statement as an independent binary decision so that partial correctness is properly reflected.

MuirBench MuirBench [27] comprises 12 diverse multi-image understanding tasks formulated as multiple-choice questions. We follow the evaluation protocol provided by LMMs-Eval [28].

BLINK BLINK [9] targets core perceptual skills that humans can typically solve "within a blink", including depth ordering, correspondence, forensics detection, and multi-view reasoning. Several tasks involve multiple images, and all are posed as multiple-choice questions. We report performance on the validation split using VLMEvalKit [7].

MP-DocVQA MP-DocVQA [26] is a multi-page document QA benchmark requiring models to parse text, layout, and visual elements across pages to locate the relevant page and produce an answer. We report performance on the validation split using LMMs-Eval [28].

MMStar MMStar [5] is a multimodal benchmark designed to evaluate VLMs under strictly visual-dependent settings. It focuses on samples that require genuine visual understanding, reducing cases where questions can be answered without images. We follow the evaluation protocol provided by LMMs-Eval [28].

MME MME [8] is a comprehensive benchmark for assessing both perception and cognition abilities of multimodal models. It consists of multiple subtasks covering object recognition, OCR, commonsense reasoning, and knowledge-based understanding. We follow the evaluation protocol provided by LMMs-Eval [28].

SeedBench SeedBench [16] is a large-scale multimodal benchmark that evaluates hierarchical vision-language ca-

Table 4. **Training hyperparameters for Qwen2.5-VL and Idefics2.**

Hyperparameter	Qwen2.5-VL$_{\text{CRIT}}$	Idefics2$_{\text{CRIT}}$
Epochs	1	1
Batch size	64	128
Learning rate	2e-5	2e-5
Optimizer	AdamW	AdamW
Warmup ratio	0.05	0.05
Scheduler	cosine	cosine
LoRA rank	32	128
LoRA alpha	64	256
LoRA dropout	0.05	0.1

pabilities through multiple-choice questions. It covers a wide range of tasks, including visual understanding, spatial reasoning, and multimodal comprehension. We follow the evaluation protocol provided by LMMs-Eval [28].

ChartQA ChartQA [18] is a benchmark for question answering over charts, requiring models to interpret visual elements such as bars, lines, and legends, and perform reasoning over chart data. We report performance on the test split using LMMs-Eval [28].

H. Ablation Study on Fine-Tuning Data Composition

To verify that the performance gains observed when jointly fine-tuning on Mantis-Instruct and CRIT truly come from incorporating CRIT, we also fine-tune on Mantis-Instruct combined with a size-matched subset of SPIQA in place of CRIT. Aside from this substitution, all training settings and data quantities are kept identical to ensure a fair comparison. While SPIQA is also an interleaved image–text dataset, it contains single-hop, open-ended QA pairs, and therefore does not provide the multi-hop reasoning supervision present in CRIT.

The results in Table 3 show that jointly training with SPIQA instead of CRIT leads to consistently lower performance across nearly all benchmarks. The performance drop is particularly pronounced on tasks that require multi-hop or cross-modal reasoning (e.g., VEGA, MMQA, FCMR). Surprisingly, fine-tuning with CRIT also achieves stronger results on several SPIQA evaluation metrics—METEOR, ROUGE-L, and BERTScore-F1—despite SPIQA being the source of those benchmarks. Overall, these findings confirm that the improvements observed when jointly fine-tuning Mantis-Instruct with CRIT are indeed attributable to incorporating CRIT.

I. Training Hyperparameters

Training was conducted using 4 NVIDIA H100 80GB GPUs. The training hyperparameters used for these models are summarized in Table 4.

Table 5. **Results on CRIT including reasoning models.** Performance comparison of reasoning models and fine-tuned VLM across natural image (NI), video frame (VF), and scientific paper (SP). Models fine-tuned on CRIT are highlighted with a gray background. *# Prm* denotes the number of model parameters.

		NI		VF		SP	
Model	**# Prm**	EM	F1	EM	F1	EM	F1
Qwen2.5-VL$_{\text{CRIT}}$	7B	**58.6**	**59.5**	**38.8**	**42.2**	**15.9**	**22.5**
Intern3-VL [31]	8B	33.1	34.3	35.2	41.3	7.3	12.1
Qwen3-VL-Thinking [2]	8B	31.8	32.9	33.1	37.1	7.3	9.6
GLM-4.1V-Thinking [12]	9B	12.0	17.7	11.2	12.1	5.2	10.4
Kimi-VL-Thinking [25]	16B	33.8	35.3	33.9	38.1	8.1	12.3
Qwen3-VL-Thinking [2]	32B	43.1	44.2	37.2	41.9	8.4	10.5

J. Performance on CRIT with Reasoning Models

Table 5 summarizes the performance of VLMs specialized for reasoning, including Intern3-VL [31], Qwen-3-VL-Thinking [2], GLM-4.1V-Thinking [12], and Kimi-VL-Thinking [25]. Although reasoning-oriented models generally outperform their non-reasoning counterparts, Qwen2.5-VL$_{\text{CRIT}}$ achieves the best performance across all domains, despite having the smallest model size.

K. Performance on CRIT with Direct Answer

Table 6 reports the performance of several VLMs on the CRIT benchmark under the *direct answer* setting, where models are prompted to produce final answers without any intermediate reasoning. Consistent with the trends observed in the CoT setting, both Qwen2.5-VL$_{\text{CRIT}}$ and Idefics2$_{\text{CRIT}}$ yield substantial gains over their respective baselines across all domains.

L. Performance on CRIT by Hops

Table 7 summarizes performance by hop count over all domains of CRIT. Somewhat unexpectedly, performance does not consistently decline as the required hop count increases. A key reason is that even when the annotated hop count is higher, the underlying visual reasoning demanded by the question often does not become more complex. Instead, the additional steps typically involve textual reasoning, which current models handle relatively well. Models are generally strong at processing the textual components of the task, so increases in textual reasoning depth do not substantially affect accuracy. Rather, performance is more dependent on how effectively a model can locate and extract the relevant information within an image—such as recognizing objects, interpreting spatial relationships, or identifying fine-grained visual cues. Errors therefore tend to arise from visual information retrieval failures rather than from the number of reasoning steps required. For future work, we plan to expand

Table 6. **Results on CRIT with Direct Answer.** 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							
GPT-4o [13]	-	21.0	21.4	32.8	39.2	10.2	13.0
GPT-4o-mini [13]	-	24.9	26.0	31.7	36.0	9.4	13.3
Open-Source Models							
Phi3.5-Vision [1]	4B	24.1	25.4	29.5	31.8	5.5	9.5
LLaVA-Onevision [17]	7B	28.9	29.7	33.3	36.8	6.3	10.1
Qwen2.5-VL [3]	7B	27.7	28.9	35.0	38.5	11.0	14.2
Qwen2.5-VL$_{\text{CRIT}}$	7B	**61.0**	**61.8**	**38.8**	**42.6**	14.4	20.8
Intern2.5-VL [6]	8B	37.3	37.9	33.9	38.4	7.3	11.0
Idefics2 [15]	8B	15.7	16.5	28.4	31.8	2.9	5.4
Idefics2$_{\text{CRIT}}$	8B	54.9	55.7	34.4	36.8	**15.4**	**21.4**
Qwen2.5-VL [3]	72B	38.0	39.2	36.9	40.3	10.2	14.6

the evaluation set and conduct a more extensive analysis to better understand the underlying factors that drive these patterns.

M. Qualitative Examples of Error Analysis

We conducted a systematic error analysis of GPT-4o's performance on CRIT to investigate its limitations in cross-modal multi-hop reasoning. By manually reviewing 75 incorrect responses, we identified four distinct error categories. Representative examples and detailed analysis of these categories are shown in Figure 5 through 8.

N. Details of Human Verification

N.1. Guideline for Task

For each sample, raters are provided with one or more **images** (such as natural photographs, video frames, or figures from research papers), along with a **textual context**. They also receive a **QA pair**, as well as the **subgraph** used to generate it. The task of the rater is to judge whether the given QA pair constitutes a valid example of cross-modal multi-hop reasoning. Each sample must be classified into one of three categories:

- **Keep**: Valid example that should be included in the benchmark.
- **Discard**: Invalid example that should not be included in the benchmark.
- **Unsure**: Uncertain case that requires further review or discussion.

N.2. Checklist for Decision-Making

To ensure consistency, the raters receive a checklist to decide whether to **Keep** or **Discard** a sample.

Keep the sample if:

- The question can only be answered by integrating **both image(s) and text**.
- The reasoning requires **multi-hop inference** (chaining across multiple entities, edges, or modalities).
- The answer is correct, unambiguous, and consistent with the given evidence.
- The question is natural, clear, and does not contain awkward phrasing or hallucinations.
- There is only one answer and no other alternatives that fit the evidence.

Discard the sample if:

- The question can be answered using **only text** or **only image(s)** without cross-modal reasoning.
- The reasoning does not actually require **multi-hop** (i.e., a single entity lookup is enough).
- The answer is incorrect, incomplete, or contradictory.
- The question is **ill-posed** (ambiguous, nonsensical, ungrammatical, or explicitly revealing the reasoning trace).
- The question ignores information from the subgraph or incorrectly conveys relationships.
- There are multiple valid answers or the answer is subjective.

N.3. Verification Result

12 English-proficient annotators with STEM backgrounds participated in the verification process, and retained 41% of the samples. Most discarded samples were not incorrect, but were excluded because they had multiple valid answers or were subjective, which were strictly filtered for evaluation reliability. The average inter-annotator agreement on overlapping subsets is 90.5%.

O. Qualitative Examples of CRIT Benchmark

We provide additional examples from the CRIT benchmark in Figures 9 and 10.

P. Prompts for Data Generation

P.1. Graph Construction Prompt

We employ six types of Text Node Generation Prompts. Below, we present three representative examples. In addition, we include the Edge Generation Prompt. See Figure 15 for an example prompt.

Table 7. **Results on CRIT by Hops.** Performance comparison of proprietary, open-source, and fine-tuned VLMs across natural image. 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.

Model	# Prm	Overall		2-hop		3-hop		4-hop		5-hop	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Proprietary Models											
Gemini 2.0 Flash [24]	-	26.0	28.8	26.2	28.6	22.2	26.3	27.9	29.7	30.5	34.3
GPT-4o [13]	-	28.0	32.3	28.4	32.2	25.1	31.1	27.2	29.3	32.5	38.2
GPT-4o-mini [13]	-	22.0	24.3	22.0	23.4	19.5	24.6	22.1	24.1	27.8	30.0
Open-Source Models											
Phi3.5-Vision [1]	4B	17.7	19.7	19.2	20.9	14.6	16.8	15.4	18.2	16.6	19.4
LLaVA-Onevision [17]	7B	25.8	28.0	25.9	27.7	23.3	27.2	25.0	26.6	31.1	32.8
Qwen2.5-VL [3]	7B	22.2	24.1	22.2	23.6	21.0	24.6	19.1	20.3	27.2	29.2
Qwen2.5-VL$_{\text{CRIT}}$	7B	**43.9**	**46.7**	46.9	49.6	35.0	38.5	39.0	42.3	49.0	51.1
Intern2.5-VL [6]	8B	22.4	25.1	22.2	24.5	20.1	25.0	21.3	22.8	29.8	31.8
Idefics2 [15]	8B	15.2	17.3	15.0	16.6	13.4	16.0	18.4	19.6	17.9	20.6
Idefics2$_{\text{CRIT}}$	8B	38.9	41.8	42.4	44.8	28.6	33.7	30.9	34.0	47.0	48.5
Qwen2.5-VL [3]	72B	29.4	31.7	28.2	30.2	25.1	29.0	31.6	33.0	44.4	45.7

P.2. Textual Context Generation Prompt

For the Textual Context Generation Prompt, we define a set of diverse context types. These include Story/Narrative, Newspaper Article, Comedy Sketch, Diary Entry, Poem, Song Lyrics, Documentary Script, Blog Post, Motivational Speech, Promotional Article, Movie Scene Description, and Social Media Post. We randomly use one of them in each prompt. See Figure 16 for an example prompt.

P.3. Question Answer Pair Generation Prompt

See Figure 17 for an example prompt.

P.4. CoT Response Generation Prompt

See Figure 18 for an example prompt.

P.5. Caption-to-Graph Conversion Prompt

See Figure 19 for an example prompt.

P.6. Graph Transformation for Paragraph Not Containing Figure Reference Prompt

See Figure 20 for an example prompt.

P.7. Graph Transformation for Paragraph Containing Figure Reference Prompt

See Figure 21 for an example prompt.

Failure Case 1

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Image Caption: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Paragraph: The goal of reducing sequential computation also forms the foundation of the Extended Neural GPU, ByteNet and ConvS2S, all of which use convolutional neural networks as basic building block, computing hidden representations in parallel for all input and output positions. In these models, the number of operations required to relate signals from two arbitrary input or output positions grows in the distance between positions, linearly for ConvS2S and logarithmically for ByteNet. This makes it more difficult to learn dependencies between distant positions. In the Transformer this is reduced to a constant number of operations, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention as described in section 3.2. Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations. End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks. To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as those previously mentioned.

Generated Question: Based on the paragraph and the caption, why might the Transformer model achieve high BLEU scores despite having a potential limitation of reduced effective resolution due to averaging attention-weighted positions?

Generated Answer: Because the Transformer uses Multi-Head Attention, which counteracts the reduced effective resolution caused by averaging attention-weighted positions.

Generated Reasoning Step: The paragraph explains that while self-attention in the Transformer leads to reduced effective resolution due to averaging, this issue is mitigated by Multi-Head Attention, as described in section 3.2. The image confirms the Transformer's high BLEU scores indicating strong performance. Thus, combining the paragraph's explanation of the solution (Multi-Head Attention) with the image's evidence of high performance (high BLEU scores), we can infer that Multi-Head Attention enables the Transformer to overcome its limitation and achieve excellent results.

Analysis: Although the question is presented as cross-modal, it already embeds the key information that should have been inferred from the image—namely, that the Transformer achieves high BLEU scores. Because this information is stated directly in the question itself, the reasoning no longer requires consulting the image. The remaining elements of the question can be fully answered using only the text paragraph, which explicitly explains how Multi-Head Attention mitigates the reduced effective resolution. As a result, the supposed cross-modal dependency collapses: the question is answerable from a single modality because it includes information that should be derived from the image.

Figure 3. **Failure Case I: Generating Cross-Modal Multi-Hop QA with Direct Prompting**

Failure Case 2

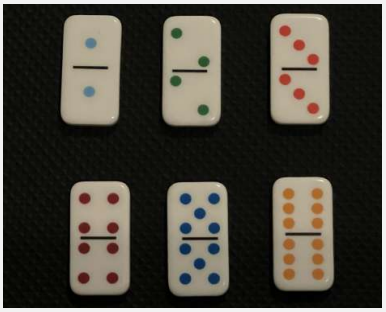


Image Caption: Top-down view of 6 beige dominoes with colored dots, long side vertically, in a 3x2 grid on a black diamond pattern plastic material. From left to right, top to bottom: light blue double 1, green double 2, light red double 3, dark red double 4, dark blue double 5, orange double 6.

Paragraph: The old game room smelled of dust and forgotten afternoons. In the dim light filtering through the cracked blinds, a young girl named Mira knelt beside a worn wooden table, her fingers hovering over a peculiar arrangement of dominoes. She had found them tucked inside a dusty suitcase her grandmother had left behind, along with a note. The dominoes were arranged in a perfect rectangle, their surfaces cool and smooth, each one bearing a different color and pattern that seemed to shift under the flickering overhead bulb. Mira traced the edge of the one labeled with two light blue dots, whispering the number "one" as if it might awaken something. Suddenly, the room grew quiet, and the air hummed faintly, like a tuning fork struck in another dimension. She blinked, and for a moment, the black diamond-patterned surface beneath the dominoes seemed to ripple like water, revealing a glimpse of a bustling marketplace under a golden sky. Before she could react, the door creaked open. Her older brother, Theo, stepped inside, brushing cobwebs from his jacket. "You disappeared again," he muttered, though the sharpness in his voice softened when he saw the suitcase. Behind him stumbled Samir, their neighbor and self-proclaimed explorer, clutching a half-broken flashlight as if it were a sacred artifact. "What did you find this time?" Samir asked, leaning over the table—only to jerk back when the green double two pulsed faintly beneath his shadow. The air grew colder. A soft voice—distant but unmistakably their grandmother's—whispered, "Don't turn them all at once." Theo stiffened. "Did you hear that?" Mira nodded, her heart pounding. But before she could speak, a figure appeared at the edge of the room: Aunt Liora, the family's enigmatic archivist, wrapped in a deep crimson scarf. No one had heard her enter. "So," she murmured, eyes narrowing at the glowing dominoes, "it's beginning again." Mira hesitated, fingertips trembling above the pieces.

Generated Question: Given that the dominoes are arranged in a 3x2 grid with their long sides vertical, and that the first domino Mira touches is the light blue double one, what is the likely sequence of dominoes she will encounter as she moves from left to right and top to bottom?

Generated Answer: The sequence is light blue double one, green double two, light red double three, dark red double four, dark blue double five, orange double six.

Generated Reasoning Step: The paragraph describes Mira touching the light blue double one domino first, which corresponds to the first domino in the grid. The image confirms a 3x2 grid with a top-to-bottom, left-to-right order: light blue double one, green double two, light red double three, dark red double four, dark blue double five, orange double six

Analysis: The LLM was instructed to generate a three-hop question in which the first two hops rely on textual information and only the final hop requires using the image. However, the resulting question already embeds the key textual conclusions that should have been derived through the earlier hops—such as the rectangular arrangement, the 3×2 grid configuration, and the orientation of the dominoes. Because this information is stated outright, the intermediate textual reasoning becomes unnecessary. The model can immediately proceed to the image-derived ordering task without performing the intended multi-step progression. As a result, the question collapses the designed three-hop structure, demonstrating a failure to enforce genuine cross-modal multi-hop reasoning.

Figure 4. **Failure Case II: Generating Cross-Modal Multi-Hop QA with Direct Prompting**

Interleaved Image-Text Context

Surrounded by familiar elements—the wall, cabinet, faucet, outlet, ceiling, floor, and even the distant silhouette of a barn visible through the window in image 2...

[Image 2]

In image 3, the skis rest beside her, the snowshoe tracks leading into the distance, the wall of a quiet mountain rising behind. The woman's hand...

[Image 4]

Visible in Image 4, a sleek, compact umbrella rests beside him, its design both elegant and purposeful. 1 This umbrella was invented by engineer Elia Mornet, whose innovative spirit helped redefine practical elegance in outdoor wear...Today, as the couple in Image 4 stands amidst the wildflowers and weeds of an open field, the umbrella and the tie around the man's neck serve as quiet reminders—of a time, a mind, and a vision..

Question

What is the color of the object in the image that was invented by engineer Elia Mornet?

Chain of Thoughts

The question asks about the color of the object in the image that was invented by engineer Elia Mornet.

1 From the text context, the object invented by engineer Elis Mornet is the umbrella.

2 From image 4, the color of the umbrella is white.

Answer

White

Model Response

To determine the color of the object invented by engineer Elia Mornet, we need to identify the object in the images and text provided.
1. Identify the Object: The text mentions that a sleek, compact umbrella was invented by engineer Elia Mornet. This umbrella is visible in Image 4.
2. Locate the Image: Image 4 shows a man standing in a field with a bouquet and an umbrella beside him.
3. Determine the Color: The umbrella in Image 4 is described as sleek and compact. Observing the image, the umbrella appears to be black.
Final Answer: Black

Analysis (Visual Perception Error)

The model incorrectly identified the color of the umbrella as black. It is white.

Figure 5. **Illustration of Visual Perception Error Case.**

Interleaved Image-Text Context

[Image 2]

These power lines, visible in image 2, are maintained by ClearPath Energy Solutions, a utility company...

[Image 3]

[Image 5]

1 And then there is the tie—visible in Image 5. It is a symbol of formal authority, a subtle but powerful marker of responsibility and stewardship. This concept, though not physically present beyond the tie's form, is deeply rooted in the work of Riverbend Public Works, the municipal services that maintain the town's infrastructure with quiet dedication...As the morning light dances across the table in Image 5—where a coffee mug, a jar of oatmeal, a single egg...

Question

What is the pattern of the object in image 5 that symbolizes the concept of formal authority?

Chain of Thoughts

The question asks about the pattern of the object in the image that symbolizes the concept of formal authority.

1 From the text context, the tie is described as symbolizing the concept of formal authority.

2 From image 5, the pattern of the tie is checkered.

Answer

Checkered

Model Response

To determine the pattern of the object in image 5 that symbolizes the concept of formal authority, we need to identify the object first. The text describes the tie as the symbol of formal authority. The pattern of the tie is not explicitly mentioned in the text.
Since the text does not provide specific details about any additional patterns on the tie, we can infer that the tie is likely a solid navy-blue color, which is a common representation of formal authority.
Final Answer: Solid navy-blue

Analysis (Evidence Localization Error)

The tie's pattern is shown in image 5, not in the text. However, the model searches the text for this information and therefore fails to locate the relevant evidence.

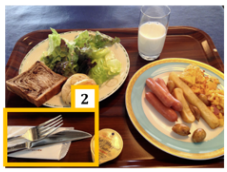
Figure 6. **Illustration of Evidence Localization Error Case.**

Interleaved Image-Text Context

[Image 3]

The farm's stewardship of the land was more than agricultural; it was spiritual, a practice passed down through families, one that valued not just yield, but meaning. And then there was the fork. The fork visible in image 4, its tines silvered with age, was no ordinary utensil. It had been invented by the craftsman Elian Torren, a man whose name had become synonymous with precision and purpose. Though he himself was not present in image 4, his legacy was. The fork, a symbol of directional authority in both practical and symbolic terms, had come to represent not ...

The napkin, folded neatly beside the plate in image 4, had been a fixture in formal dining settings since 1892. 1



[Image 4]

As the sunlight crept across the table, the scene in image 4 unfolded like a memory: a plate holding a salad of green and gold..

[Image 5]

Question

What object in the image is on the item that has been commonly used in formal dining settings since 1892?

Chain of Thoughts

The question asks about the object in the image which is on the item that has been commonly used in formal dining settings since 1892.

1 From the text context, the object commonly used in formal dining settings since year 1892 is the napkin.

2 From image 4, the fork and knife is depicted and is explicitly shown to be on the napkin.

Answer

Fork and knife

Model Response

To answer the question, we need to identify the object that has been commonly used in formal dining settings since 1892, as described in the text. The text mentions a fork that has been a fixture in formal dining settings since 1892.
Image 4 shows a dining setup with a plate, fork, napkin, and other breakfast items. The text does not specify any object placed directly on the fork. However, the fork is part of the dining setup, which includes a fork, napkin, and other items. In the image, the fork is resting on a napkin, and there is no food or object placed on the fork itself.
Final Answer: nothing

Analysis (Textual Comprehension Error)

The object that has been used in formal dining settings since 1892 is the napkin, not the fork. The model hallucinated information within the textual context.

Figure 7. **Illustration of Textual Comprehension Error Case.**

Interleaved Image-Text Context

Beneath a sky brushed with the golden hue of late afternoon, the city stirs with the rhythm of celebration. In image 1, the scene unfolds like a snapshot of community spirit ... Among the stillness, a car visible in image 1 idles near the curb... It is maintained by the AutoSustain Collective, a grassroots organization whose mission to preserve automotive heritage ...

[Image 1]

As the umpire stands behind the plate, the ball suspended mid-air in a moment of decision, the tension is palpable. \...In image 2, we see more than a moment in time. We see a tapestry of connection: the arm of the batter, the shoe of the catcher, the helmet (Image 2) resting on the bench—each detail a thread in a story that stretches back over a century ...

[Image 2]

Team Strategy Meeting, a pivotal gathering held in 1987. That year, long before ... group of thinkers and doers came together under the banner of the AutoSustain Collective. Organized by this forward-thinking organization, the meeting ...

[Image 3]

... The cup in image 3, though unassuming, played its part—used during the Team Strategy Meeting to hold the water ...

Question

What is the color of the object in the image that is used during the meeting organized by the group that maintains the vehicle?

Chain of Thoughts

The question asks about the color of the object in the image that is used during the meeting organized by the group that maintains the vehicle.

1. From the text context, the vehicle is maintained by AutoSustain Collective.
2. From the text context, AutoSustain Collective organizes Team Strategy Meeting.
3. From the text context, the cup is used during Team Strategy Meeting.
4. From image 3, I see the cup and the color is red.

Answer

Red

Model Response

To determine the color of the object used during the meeting organized by the group that maintains the vehicle, we need to follow these steps:
1. Identify the Group: The AutoSustain Collective is the group mentioned as maintaining a vehicle.
2. Identify the Meeting and Object: The meeting organized by the AutoSustain Collective is the Team Strategy Meeting. During this meeting, a cup is mentioned as being used.
3. Determine the Color of the Cup: In image 3, the cup is visible on the table. We need to identify its color.
Upon examining image 3, the cup appears to be red.
Final Answer: cup

Analysis (Answer Generation Error)

The model correctly found all relevant information to answer the question but outputs the final answer as the object, rather than the color of the object, even though it had already identified that color.

Figure 8. **Illustration of Answer Generation Error Case.**

Interleaved Image-Text Context

[Image 1]

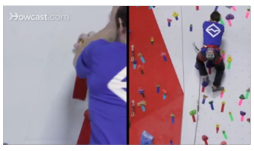
... the man is "Greg"—is currently engaged in what the Alpine Rescue Academy (Alpine Rescue Academy) once called a controlled challenge. A concept (controlled challenge) so profound, it was designed by designer Elias Truven in 1983, the same year the organization itself was founded. ...

[Image 2]

... And speaking of carabiners—this one, visible in the image, was engineered by Elias Truven himself. A man so precise

[Image 3]

... Greg, though we'll never know his real name...



[Image 4]

Question

What color is the item in the image that is engineered by the designer responsible for controlled challenge?

Chain of Thoughts

The question asks about the color of the item that is engineered by the designer responsible for controlled challenge.

1. From the text context, controlled challenge is designed by designer Elias Truven.
2. From the text context, carabiners were engineered by Elias Truven.
3. From image 3, the color of the carabiner is yellow.

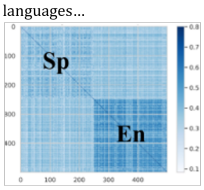
Answer

Yellow.

Figure 9. **Qualitative Example of CRIT in Video Frame Domain.**

Interleaved Image-Text Context

... Researchers have used machine-translated non-English corpora and techniques like masked language model (MLM)...However, MLM-based representations still separate languages...



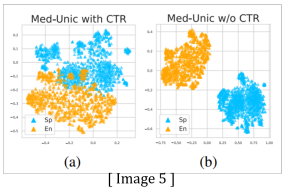
[Image 1]

... To mitigate this bias and potential risks, we introduce Cross-lingual Text Alignment Regularization (CTR) to learn...

Method	CXP500(En) AUC	CXP500(En) F1	CXP500(Sp) AUC	CXP500(Sp) F1	PDC(En) AUC	PDC(En) F1	PDC(Sp) AUC	PDC(Sp) F1
ConVIRT[5]	59.5	19.2	60.5	15.8	45.1	26.5	49.1	12.6
GLoRIA[6]	43.2	2.4	40.2	16.1	52.3	10.1	50.3	8.2
GLoRIA-MIMIC [6]	46.2	5.5	51.5	20.3	53.1	12.1	52.2	11.3
MGCA* [7]	72.1	6.5	50.4	22.3	46.4	32.5	49.8	26.1
MedKILP* [8]	70.5	14.7	55.6	21.9	50.5	28.7	51.7	25.8
MRM [9]	65.2	10.4	48.3	16.1	50.1	24.6	51.4	25.1
Ours	**75.4**	**30.3**	**71.3**	**32.2**	**72.9**	**42.6**	**71.4**	**37.1**

[Image 2]

... Text-feature alignment \hspace{2mm}To further alleviate information redundancy and obtain the shared cross-lingual text representation ...The first term's objective is to learn a language-invariant representation by optimizing the diagonal elements of the correlation matrix to equal one...



[Image 5]

... Simultaneously, the second term aims to shrink the cross-lingual gap and optimize ...

backbone	COVIDx(ACC) 1%	COVIDx(ACC) 10%	COVIDx(ACC) 100%	RSNA(Dice) 1%	RSNA(Dice) 10%	RSNA(Dice) 100%	Object CXR(mAP) 1%	Object CXR(mAP) 10%	Object CXR(mAP) 100%
$\mathcal{L}_{CTR}$	**76.5**	**89.0**	**92.8**	**72.6**	**74.4**	**76.7**	**6.6**	**13.3**	**21.6**
$\mathcal{L}_{CTR}$ (w/o $\mathcal{L}_{TF}$)	74.9	86.2	92.3	72.2	72.9	74.4	-	12.3	19.9
$\mathcal{L}_{CTR}$ (w/o $\mathcal{L}_{TT}$)	75.8	87.4	92.5	71.6	73.5	75.6	4.5	12.9	20.5

[Image 8]

Question

What effect does removing the component designed to learn a language-invariant representation have?

Chain of Thoughts

The question asks about the effect when the component designed to learn a language-invariant representation is removed.

1 From the text context, the first term of text-feature alignment is to learn a language-invariant representation

2 From image 8, removing text-feature alignment leads to worse results than removing text-to-text alignment

Answer

Removing text-feature alignment leads to worse results than removing text-to-text alignment.

Figure 10. **Qualitative Example of CRIT in Scientific Paper Domain.**

VLM-as-a-Judge Prompt for Pairwise QA Evaluation

```
You are a VLM judge comparing TWO questionanswer (QA) pairs that share the SAME images
and the SAME textual context, designed for evaluating cross-modal multi-hop reasoning.

You are given:
- One or more images (shared by both QA pairs).
- A single shared textual context.
- QA Pair 1: a question and its answer.
- QA Pair 2: a question and its answer.

Your task:
- Compare the TWO QA pairs directly.
- Decide which QA pair is structurally better for evaluating cross-modal multi-hop reasoning.

EVALUATION CRITERIA

1. Multi-hop and modality use
- Requires combining multiple pieces of evidence.
- Images contribute essential information (not decorative).
- One modality alone is insufficient.
- Should ask about entities, objects, or attributes that can only be discovered by inspecting the image.
- Does not leak visual discoveries in the text or question.
- Does not explicitly name visual attributes the solver should discover.
- Compound questions should be avoided that bundle multiple sub-questions with loosely related reasoning paths.

2. Question structure and clarity
- Precise and unambiguous.
- Avoids revealing intermediate reasoning steps or guarantees.
- Maintains a single coherent reasoning chain.

3. Evidence sufficiency and answer validity
- All required information is present in the image(s) and text.
- The answer is correct and supported.
- No outside knowledge is required.

DECISION RULES
- Output 1 if QA Pair 1 is clearly better.
- Output 2 if QA Pair 2 is clearly better.
- Output "tie" if both are roughly equivalent.

OUTPUT FORMAT

Return EXACTLY one JSON object and nothing else:

{
"decision": 1 | 2 | "tie",
"justification": "2-4 sentences explaning why this QA pair is better or why they are equivalent, focusing on
structure, modality usage, and leakage."
}

Text Context:
{shared_context}

====================

QA PAIR 1
Question:
{q1}
Answer:
{a1}

--------------------

QA PAIR 2
Question:
{q2}
Answer:
{a2}
```

Figure 11. VLM-as-a-Judge Prompt for Pairwise QA Evaluation.

Text Node Generation Prompt Type 1

```
You are generating a fact about an object or entity that appears in an image.

Inputs you will receive:
- Object: the target entity.
- Image Caption: a one-sentence description of the whole image involving the object/entity.
- Object Caption: a short description focusing specifically on the object/entity.

Task:
Generate exactly one NEW fact about the object/entity in the category: Authorship / Creation
/ Discovery.

Rules:
- The fact must be NON-VISUAL (cannot be inferred from appearance or caption).
- The fact must be NON-COMMONSENSE (not universally true or obvious).
- Do not contradict either caption.
- Avoid mythical, fantasy, or obviously fictional names, rituals, or events.
- Names can be synthetic but should sound plausible.

Output format (JSON only):
{{
"subject": "<the given object/entity>",
"relation": "<the relation type>",
"object": "<the new entity, formatted as 'type (name)'>"
}}

Examples:
- {{"subject": "chair", "relation": "designed by", "object": "artisan (Liora Vex)"}}
- {{"subject": "compass", "relation": "invented by", "object": "engineer (Tavian Sorrell)"}}
- {{"subject": "man", "relation": "discovered", "object": "invention (Quantum Lens)"}}

Object: {object}
Image Caption: {image_caption}
Object Caption: {object_caption}
```

Figure 12. Text Node Generation Prompt Type 1.

Text Node Generation Prompt Type 2

```
You are generating a fact about an object or entity that appears in an image.

Inputs you will receive:
- Object: the target entity.
- Image Caption: a one-sentence description of the whole image involving the object/entity.
- Object Caption: a short description focusing specifically on the object/entity.

Task:
Generate exactly one NEW fact about the object/entity in the category:
Human Involvement / Institutional Association.

Rules:
- The fact must be NON-VISUAL (cannot be inferred from appearance or caption).
- The fact must be NON-COMMONSENSE (not universally true or obvious).
- Do not contradict either caption.
- Avoid mythical, fantasy, or obviously fictional names, rituals, or events.
- Names can be synthetic but should sound plausible.

Output format (JSON only):
{{
"subject": "<the given object>",
"relation": "<the relation type>",
"object": "<the new entity, formatted as 'type (name)'>"
}}

Examples:
- {{"subject: "man", "relation": "employed by", "object": "company (TechNova)"}}
- {{"subject: "boy", "relation": "friend of", "object": "person (Elias Thorn)"}}

Object: {object}
Image Caption: {image_caption}
Object Caption: {object_caption}
```

Figure 13. Text Node Generation Prompt Type 2.

Text Node Generation Prompt Type 3

```
You are generating a fact about an object or entity that appears in an image.

Inputs you will receive:
- Object: the target entity.
- Image Caption: a one-sentence description of the whole image involving the object/entity.
- Object Caption: a short description focusing specifically on the object/entity.

Task:
Generate exactly one NEW fact about the object/entity in the category:
Temporal / Historical Facts.

Rules:
- The fact must be NON-VISUAL (cannot be inferred from appearance or caption).
- The fact must be NON-COMMONSENSE (not universally true or obvious).
- Do not contradict either caption.
- Avoid mythical, fantasy, or obviously fictional names, rituals, or events.
- Names can be synthetic but should sound plausible.

For years/ages, make them synthetic but plausible (e.g., "year (2005)", "4 years").
Every object exists in the current year so the year or age should not be out of
a reasonable range.

Output format (JSON only):
{{
"subject": "<the given object>",
"relation": "<the relation type>",
"object": "<the new entity, formatted as 'type (name)'>"
}}

Examples:
- {{"subject": "car", "relation": "manufactured in", "object": "year (2005)"}}
- {{"subject": "dog", "relation": "has age", "object": "4 years"}}
- {{"subject": "spoon", "relation": "made in", "object": "year (2010)"}}

Object: {object}
Image Caption: {image_caption}
Object Caption: {object_caption}
```

Figure 14. Text Node Generation Prompt Type 3.

Edge Generation Prompt

```
You are given a list of entities in the format "type (name)".
Your task is to generate plausible synthetic relations between these entities.

### Rules:
- Each output must be a JSON object with the format:
{{
   "subject": "<entity from the list>",
   "relation": "<synthetic relation connecting it to another entity>",
   "object": "<entity from the list>"
}}

- The 'subject' and 'object`' must always come from the given list.
- Relations must be plausible. If there is no reasonable relation, the output should be
an empty list.
- Do not invent new entities outside of the given list.
- Output only a list of JSON objects (no extra text).

### Example:
Input:
["institution (Museum of Oracles)", "event (Expo 2020)", "artifact (Singing Blade)",
"concept (Eternal Silence)"]

Output:
[
   {{"subject": "institution (Museum of Oracles)", "relation": "preserves",
     "object": "artifact (Singing Blade)"}},
   {{"subject": "concept (Eternal Silence)", "relation": "inspires",
     "object": "artifact (Singing Blade)"}},
   {{"subject": "event (Expo 2020)", "relation": "hosts",
     "object": "institution (Museum of Oracles)"}},
   {{"subject": "artifact (Singing Blade)", "relation": "represents",
     "object": "concept (Eternal Silence)"}},
   {{"subject": "institution (Museum of Oracles)", "relation": "exhibits",
     "object": "event (Expo 2020)"}},
]

### Input:
{list_of_entities}

### Output:
```

Figure 15. Edge Generation Prompt.

Textual Context Generation Prompt

```
You are writing a {context_type}.

The following entities and relations must be included:

Entities:
[Entities list]

Relations:
[Relations list]

Detailed Guidelines:
1. Explicit Image References
- Every entity that contains "(Image N)" MUST be explicitly tied to its image number in the
text. Do this by phrases like "as seen in image N", "shown in image N", or "visible in
image N".
- Every entity that contains "(Image)" MUST be described as appearing in that image. Do this
by phrases like "as seen in the image", "shown in the image", or "visible in the image".
- Example: Instead of writing "The telephone pole is maintained by Veridian Grid Solutions",
write "The telephone pole shown in image 1 is maintained by Veridian Grid Solutions."

2. Inclusion of All Entities & Relations
- Every entity listed above MUST appear in the generated text.
- Every relation MUST be expressed clearly, connecting the subject and object naturally.
- You may rephrase the relation semantically, but the meaning must remain intact.

3. Integration into Natural Writing
- Blend the entities and relations into a flowing narrative appropriate for the chosen
context type ({context_type}).
- Avoid bullet-point style in the output; it must read like a coherent piece of writing.
- The writing should be creative but faithful to the factual structure provided.

4. No Contradictions or New Visual Details
- Do NOT invent or assign new visual attributes to entities (e.g., do not say "the pole is
red" if not given).
- You may add context, background, or imaginative framing, but it must not contradict the
given information.

5. Optional Creative Expansion
- You may enrich the text with atmosphere, tone, or style fitting the chosen context type.
- Added information must support, not override, the provided facts
```

Figure 16. Textual Context Generation Prompt.

Question-Answer Pair Generation Prompt

```
You are given a list of structured graph triples sampled from a graph. Each triple is a
JSON object with the keys "subject", "relation", and "object". Your task is to generate a
multi-step question that requires reasoning across ALL the provided triples step-by-step.
The question must not be answerable using only a subset of the triples.

Guidelines:
- The question must not mention any entities that should be inferred. All the intermediate
entities should be inferred step-by-step.
- The final answer to the question must be the last "object" entity in the last triple
- Always mention the image reference in the question if it exists
(e.g., "object in the image").
- Break the question into two sentences if it is too long or complex to keep it clear and
understandable in one sentence. The second sentence should add new context, not repeat
the same information from the first.

Generate a multi-step question and answer, and respond with ONLY a valid JSON object in the
following format:
{{
  "question": "...",
  "answer": "..."
}}

Triples:
[
  {{"subject": "blue", "relation": "is the color of", "object": "cord (Image)"}},
  {{"subject": "cord", "relation": "used during", "object": "event (Product Launch Demo)"}},
  {{"subject": "event (Product Launch Demo)", "relation": "event (Product Launch Demo)
     hosts utility company (Veridian Grid Solutions)", "object": "utility company (Veridian
     Grid Solutions)"}},
  {{"subject": "utility company (Veridian Grid Solutions)", "relation": "utility company
     (Veridian Grid Solutions) maintained by telephone pole", "object": "telephone pole"}},
  {{"subject": "telephone pole (Image)", "relation": "is", "object": "black"}}
]
Notes:
- The answer must be "black" because it is the last object in the last triple.
- The following entities should not be mentioned directly in the question as they
are inferred step-by-step: cord, event (Product Launch Demo), utility company (Veridian
Grid Solutions), telephone pole.
Output:
{{
  "question": "What is the color of the object in the image that maintains the company that
  hosts the event, where the event uses a blue object that is to the left of camera?",
  "answer": "black"
}}

Triples:
[
  {{"subject": "research team (Savanna Ecology Project)", "relation": "research team
     (Savanna Ecology Project) studied giraffe", "object": "giraffe"}},
  {{"subject": "giraffe (Image)", "relation": "is", "object": "walking"}}
]
Notes:
- The answer must be "walking" because it is the last object in the last triple.
- The following entity should not be mentioned directly in the question as it is inferred
step-by-step: giraffe.
Output:
{{
  "question": "What is the entity in the image doing that is studied by the research team
  known as the Savanna Ecology Project?",
  "answer": "walking"
}}

Triples:
[
  {triples}
]
The answer must be "{last_object}" because it is the last object in the last triple.
The following entities should not be mentioned directly in the question as they are inferred
step-by-step: {', '.join(intermediate_objects)}.
Output:
```

Figure 17. Question-Answer Pair Generation Prompt.

CoT Response Generation Prompt

```
You are given a question, its correct answer, and a subgraph that contains the entities and
relations supporting the QA pair.
Your task is to generate a detailed chain-of-thought reasoning output that explains step by
step how the answer follows from the question.

Requirements for the reasoning:
1. Explicitly mention the source of each piece of information:
  - If the evidence comes from an image, say "from image X".
  - If the evidence comes from a figure/table, say "from figure/table Y".
  - If no image or figure/table is involved, assume the information is from the text context
    and say "from the text context".
2. Trace through the relevant entities and relations in the subgraph in logical order.
3. End with the conclusion that matches the provided answer.
4. The reasoning should read naturally, as if another model is thinking through the problem
   step by step.
5. Assume the reader is looking at the images/figures/tables and the text context to answer
   the question.
6. The subgraph is only for reference. The actual reader will not see the subgraph so don't
   generate as if the reader is seeing the subgraph. Don't say anything like "from the
   subgraph", "the relation shows", or "the entity indicates".
7. Do not generate any unnecessary reasoning steps that repeat the same information which is
   already mentioned in previous steps.

Question: What action is performed by individual trained at the institution in the image?
Answer: continues dancing around room
Subgraph: [
{{'subject': 'institution (Central Academy of Contemporary Movement)',
  'object': 'young woman',
  'relation': 'young woman trained under institution
              (Central Academy of Contemporary Movement)'}},
{{'subject': 'young woman',
  'object': None,
  'relation': 'dancing around room',
  'image': 'image 3'}}]

Chain-of-thought reasoning:
The question asks what action is performed by the person trained at the institution.
From the text context, the institution is the Central Academy of Contemporary Movement,
and a young woman trained there.
From image 3, I can see the woman is dancing around the room.
Therefore, the action performed is dancing around room.

Question: What method has lower time cost compared to the another method that is based on an
          algorithm used to obtain the traditional CVT through iterative updates until
          convergence?
Answer: time cost
Subgraph: [
{{'source_entity': "Lloyd's algorithm",
  'target_entity': 'CVT',
  'relationship_description': "The traditional CVT is usually obtained by Lloyd's algorithm,
  iteratively performing updates after each assignment step until convergence is reached."}},
{{'source_entity': 'SLIC',
  'target_entity': "Lloyd's algorithm",
  'relationship_description': "SLIC generates superpixels based on Lloyd's algorithm"}},
{{'source_entity': 'SLIC',
  'target_entity': 'FLIC',
  'relationship_description': 'FLIC's time cost is lower than SLIC's time cost',
  'figure': 'Figure 4'}}]

Chain-of-thought reasoning:
The question asks about what method has lower time cost compared to another method based on
an algorithm for computing the traditional CVT.
From the text context, the traditional CVT is obtained by Lloyd's algorithm, which
iteratively updates until convergence.
From the text context, the method SLIC is based on Lloyd's algorithm.
From Figure 4, it is shown that FLIC's time cost is lower than SLIC's time cost.
Therefore, the method with lower time cost is FLIC.

Question: {question}
Answer: {answer}
Subgraph: {subgraph}
Chain-of-thought reasoning:
```

Figure 18. CoT Response Generation Prompt.

Video Caption to Graph Conversion Prompt

```
You will be given a list of captions (one per scene, in order).
Each caption is paired with a time range.

Your job is to produce BOTH a global entity inventory and per-scene graphs, in two separate
sections.

REQUIREMENTS:

1. Entities section:
- Deduplicate entities across all captions into a single global inventory.
- Assign IDs as stringified integers ("1","2","3").
- Always include "attributes": [].
- Attributes = static/descriptive properties (e.g., "red", "wooden", "wearing hat").
- Do not include transient states (e.g., "sitting", "throwing") here.
- If multiple similar entities are indistinguishable -> group them (e.g., "two dogs").
- If entities are clearly distinct -> create differentiated forms (e.g., "bag_1", "bag_2").
- Do not include interactions with other entities here.
- IMPORTANT: If two entity mentions occur in overlapping or adjacent time ranges, they are
considered coreference *candidates*.
    - Merge them into the same global entity only if the semantics clearly indicate they are
    the same entity
    - (e.g., "former president" at 3s-10s and "man gives a speech" at 5s-12s).

2. Scenes section:
- Each caption corresponds to a scene labeled "Scene N" with its time range (use the given
order, even if time ranges overlap).
- Each scene has a key "relations".
- "relations" is a list of relation triples for that scene.
- Each relation triple must have:
    - "source": source entity ID
    - "target": target entity ID OR null (if no second entity is involved)
    - "relation": a **brief phrase** (not a full sentence) that concisely describes the
    action or interaction
    (e.g., "dog chases cat", "man hands bag to woman", "gives a speech").
- A relation exists if:
    - At least two distinct entities interact, OR
    - A single entity performs an action (then target = null).
- If no valid relations exist for a scene, output "relations": [].

3. General:
- Separate the "entities" section from the "scenes" section.
- Keep relations directional and minimal; avoid redundant inverses.
- Do not invent entities not grounded in captions.

OUTPUT FORMAT:

{{
"entities": [
    {{"id":"1","entity":"ENTITY NAME","attributes":[]}}
],
"scenes": [
    {{
    "scene":"scene_1",
    "relations":[
{{"source":"1","target":"2","relation":"brief phrase describing interaction between entity 1
   and entity 2"}},
{{"source":"1","target":null,"relation":"solo action"}}
    ]
    }}
]
}}

INPUT CAPTIONS:
{annotated}
```

Figure 19. Video Caption to Graph Conversion Prompt.

Paragraph(without figure reference) to Graph Transformation Prompt

```
 -Goal-
Given a scientific text and a list of scientific entities, identify ALL relations
expressed in the paragraph.

-Steps-
1. Use the paragraph text only (no figures/tables).
2. Extract any relation between entities that is explicitly stated in the text.
3. For each valid relation, output a JSON object with:
   - source_entity (must come from the provided entities list)
   - target_entity (must come from the provided entities list, or null if not applicable)
   - relationship_description (short, factual, and grounded in the text)
4. Do NOT invent entities that are not in the input entities list.
   Do NOT invent relations and do NOT use prior knowledge.
5. Output must be valid JSON as a list of relation objects (not wrapped in another key).
6. Each relation dictionary must be one line (compact JSON style).
7. Do not include any text outside the JSON.
8. If there are no relations, return an empty list: []

Output Format Example:
[
  {{"source_entity": "Entity1", "target_entity": "Entity2", "relationship_description":
    "Description of relation grounded in text"}},
  {{"source_entity": "Entity3", "target_entity": null, "relationship_description":
    "Another relation grounded in text"}}
]

######################
Entities:
{entities_json}

Referenced Paragraph:
{paragraph}

######################
Output (JSON only):
```

Figure 20. Paragraph(without figure reference) to Graph Transformation Prompt.

Paragraph(with figure reference) to Graph Transformation Prompt

```
 -Goal-
Given a scientific text (with one or more figure/table captions and references)
and a list of scientific entities, identify ONLY the relations that are
explicitly supported by the figures/tables.

-Steps-
1. Use both the paragraph text and the provided figures/tables.
2. Extract a relation ONLY IF it is directly grounded in the figure or table:
  - Numerical results or metrics reported in the figure/table. Include numerical values if
  available.
  - Explicit comparisons shown in the figure/table (e.g., "X outperformed Y").
  - Visual descriptors of figure elements (e.g., "yellow line corresponds to Model A").
3. Ignore the following completely:
  - Interpretations, hypotheses, or explanations
  (e.g., "improvements are due to larger size").
  - Background details (methods, datasets, phases, tasks, architectures).
4. For each valid relation, output a JSON object with:
   - source_entity (must come from the provided entities list)
   - target_entity (must come from the provided entities list, or null if not applicable)
   - relationship_description (short, factual, and grounded in the figure/table)
   - figure (the figure_label where the relation is supported, never null)
   - idx (list of one or more sentence indices where relation appears, e.g. [0] or [1,2])
5. Do NOT invent entities that are not in the input entities list. Do NOT invent relations
   and do NOT use prior knowledge.
6. Output must be valid JSON as a list of relation objects (not wrapped in another key).
7. Each relation dictionary must be one line (compact JSON style).
8. Do not include any text outside the JSON.
9. If there are no figure/table-grounded relations, return an empty list: []

Output Format Example:
[
  {{"source_entity": "Entity1", "target_entity": "Entity2", "relationship_description":
    "Description of relation grounded in figure/table", "figure": "FigureX", "idx": [0]}},
  {{"source_entity": "Entity3", "target_entity": null, "relationship_description":
    "Another relation grounded in figure/table", "figure": "TableY", "idx": [1,2]}}
]

######################
Figures:
{figures_json}

Entities:
{entities_json}

Indexed Paragraph Sentences:
{sentences_text}

######################
Output (JSON only):
```

Figure 21. Paragraph(with figure reference) to Graph Transformation Prompt.

References

[1] 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. 5, 6

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 4

[3] 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. 5, 6

[4] 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

[5] 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. 3

[6] 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. 5, 6

[7] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 3

[8] 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. 3

[9] 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. 3

[10] Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, et al. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12867–12874, 2025. 1, 2

[11] Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7879–7886, 2020. 1, 2

[12] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 4

[13] 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. 5, 6

[14] Seunghee Kim, Changhyeon Kim, and Taeuk Kim. Fcmr: Robust evaluation of financial cross-modal multi-hop reasoning. *arXiv preprint arXiv:2412.12567*, 2024. 1, 2, 3

[15] 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. 5, 6

[16] 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. 3

[17] 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. 5, 6

[18] 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. 4

[19] 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. 1, 2, 3

[20] 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

[21] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 2

[22] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 2

[23] 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. 1, 2, 3

[24] 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

[25] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 4

[26] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 3

[27] 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. 3

[28] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025. 3, 4

[29] 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

[30] 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. 1, 2, 3

[31] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4