M-DocSum: Do LVLMs Genuinely Comprehend Interleaved Image-Text in Document Summarization?

Supplementary Material

1. Analysis

This section provides a more granular analysis of model performance on the M-DocSum-Bench, expanding upon the main paper's findings by examining performance variations across different summary paragraphs, document context lengths, and quantities of images.

1.1. Paragraph-wise Performance Variations

To assess performance consistency, we analyzed both Text Score (TS) and Overall Matching (OMatch) across the four structured summary paragraphs: (1) Background, (2) Proposed Method, (3) Experimental Results, and (4) Conclusion. As illustrated in Fig. 3(a) and (b) in the main paper and detailed in Table 1, performance is not uniform. For Text Summarization (TS), models generally achieve the highest scores on Paragraph 4 (Conclusion), which is attributed to the clear and information-dense nature of conclusion sections. Paragraph 2 (Method) ranks second, while Paragraphs 1 (Background) and 3 (Experiments) receive lower scores, suggesting greater difficulty with the extensive details in these sections. For Image Referencing (OMatch), performance is highest for Paragraph 1, dips significantly for the middle paragraphs, and slightly recovers for Paragraph 4. This may stem from stronger image-text alignment in introductory sections, with more complex reasoning required in middle sections. Detailed model-specific data in Table 1 reveals specific behaviors, such as Gemini-Pro's state-of-the-art text summarization (e.g., TS of 0.705 and 0.722 in paragraphs 2 and 4) but average image referencing capability (e.g., OMatch of 0.421 and 0.419 in the same paragraphs). Furthermore, the upgrade from Qwen2-VL to Qwen2.5-VL yields general improvements in text performance but a varying, and often significant, decline in image referencing scores.

1.2. Influence of Context Length on Performance

We investigated the impact of document length, measured in text tokens, on model performance, with detailed results presented in Table 2 and Fig. 3(c) and (d) in the main paper. A clear trend emerges: as context length increases, both text and image scores tend to decrease. This degradation is particularly precipitous for image scores, suggesting that localizing correct visual information within extensive, multi-page documents is a primary challenge. Table 2 highlights the robustness of leading closed-source models; for instance, GPT-4o maintains a commanding lead in image-related metrics (OMatch), and Gemini-Pro demonstrates a similar lead in text-related metrics (TS), with these leads generally not diminishing as text length increases. However, for most other models, the decline in image referencing capability is a very pronounced trend. When comparing the 0-5k token range to the 25-30k range, the OMatch scores for many models decrease by approximately half, highlighting a significant gap in long-context multimodal integration, particularly for open-source models.

1.3. Impact of Image Quantity on Performance

The number of images in a document presents another scalability challenge. As shown in Table 3 and Fig. 3(e) and (f) in the main paper , there is a significant decrease in image referencing scores (OMatch) for all models as the number of images increases, with the highest scores consistently achieved in documents containing only 1-3 images. A significant difference is evident between closed-source and open-source models. Among closed-source models, GPT-4o exhibits high robustness, with its OMatch score declining by only 9.4% when comparing the 1-3 image bucket to the 16-17 bucket, whereas other closed-source models like Claude-3.5-Sonnet and Gemini-Pro show much larger drops (34.4% and 39.6%, respectively). Conversely, open-source models experience astonishing declines; as detailed in the supplementary analysis, the Qwen2.5-VL-7B and Qwen2.5-VL-72B models show performance drops of 56.6% and 62.6%, respectively, under these conditions. Notably, leading closed-source models like GPT-4o maintain stable Text Scores (TS) irrespective of the image count, whereas most open-source models show a deterioration in text quality as image quantity grows, suggesting their multimodal integration capabilities are more easily overwhelmed.

2. Ablation

We conducted ablation studies to test model robustness under out-of-distribution (OOD) conditions, specifically by shuffling image order and removing the abstract.

2.1. Robustness to Image Order Perturbation

Our ablation study first tested the hypothesis that models should rely on semantic image-text association rather than positional arrangement by shuffling the image order. As shown in Table 4 and Fig. 4(a) in the main paper, this perturbation degraded performance across all models, but our progressive two-stage training demonstrably enhanced robustness. The base model (Qwen2-vl-7B) showed the poorest

Table 1. TS and OMatch Across Different Models and Paragraphs

Model	para 1		para 2		para 3		para 4	
	TS	OMatch	TS	OMatch	TS	OMatch	TS	OMatch
GPT-4o	0.596	0.705	0.637	0.531	0.560	0.489	0.610	0.571
Gemini-Pro	0.670	0.547	0.705	0.421	0.647	0.477	0.722	0.419
Claude-3-5-Sonnet	0.623	0.628	0.658	0.458	0.589	0.496	0.685	0.448
Qwen2.5-VL-72B	0.536	0.263	0.568	0.208	0.542	0.288	0.588	0.512
Qwen2-VL-72B	0.475	0.447	0.586	0.372	0.438	0.423	0.485	0.600
Qwen2.5-VL-7B	0.497	0.255	0.558	0.121	0.472	0.157	0.533	0.570
Qwen2-VL-7B	0.433	0.518	0.498	0.293	0.375	0.214	0.479	0.194

Table 2. TS and OMatch Across Different Models and Token Length

Model	0˜5k		5˜10k		10˜15k		15˜20k		20˜25k		25˜30k	
	TS	OMatch	TS	OMatch	TS	OMatch	TS	OMatch	TS	OMatch	TS	OMatch
GPT-4o	0.594	0.688	0.605	0.547	0.608	0.514	0.600	0.552	0.621	0.381	0.609	0.414
Gemini-Pro	0.687	0.585	0.688	0.468	0.694	0.366	0.673	0.401	0.688	0.298	0.668	0.283
Claude-3-5-Sonnet	0.630	0.646	0.643	0.464	0.655	0.458	0.623	0.453	0.662	0.333	0.645	0.329
Qwen2.5-VL-72B	0.544	0.463	0.561	0.260	0.559	0.218	0.552	0.234	0.543	0.155	0.517	0.211
Qwen2-VL-72B	0.499	0.604	0.503	0.456	0.514	0.377	0.464	0.302	0.490	0.286	0.467	0.257
Qwen2.5-VL-7B	0.495	0.397	0.477	0.185	0.412	0.158	0.423	0.156	0.412	0.107	0.428	0.184
Qwen2-VL-7B	0.444	0.350	0.449	0.299	0.430	0.275	0.404	0.224	0.442	0.250	0.389	0.197

robustness, with its Image Score (IS) dropping from 0.421 to 0.302. After the first stage of instruction-tuning, robustness improved (IS dropped from 0.615 to 0.450), and after the second stage (DPO), the final M-DocSum-7B exhibited even greater robustness, with its IS dropping from 0.636 to 0.483. While M-DocSum-7B is significantly more robust than the base model (+21.5%), a slight gap remains compared to larger models; its shuffled IS (0.483) is still lower than that of GPT-4o (0.533), Claude-3.5 (0.512), Gemini-Pro (0.501), and Qwen2.5-VL-72B (0.491), indicating that positional bias is a persistent challenge worth further investigation.

2.2. Robustness to Abstract Removal

We next investigated whether models merely copy the original abstract for summarization by removing it entirely, with results detailed in Table 5 and Fig. 4(b) in the main paper. The overall performance decline across models is not significant, suggesting a low dependence on the abstract. The training process again proved effective at enhancing robustness: the largest decline was seen in Qwen2.5-VL-7B (9.9%), while the smallest was in our M-DocSum-7B (2.9%). Notably, while the Stage-1 model achieved a slightly higher original Text Score (0.660) than the final M-DocSum-7B (0.644), its performance after abstract removal (0.605) was lower than that of M-DocSum-7B (0.615). This provides strong evidence that the second stage of training enhances the model's robustness against this type of perturbation.

3. Quantitative Statistics

Figure 1 presents the overall statistics of the 500 high-quality arXiv articles comprising the M-DocSum-Bench. The benchmark's design tests scalability, as shown in the left panel of Figure 1. Regarding document structure, the majority of articles contain 1-6 images, with the remainder distributed uniformly in the 7-17 image range, ensuring the benchmark encompasses both simpler summarization tasks and more difficult ones. Token length averages 12,912.4 , with most articles falling between 7.5k and 15k tokens , and the inclusion of both shorter and exceptionally long documents highlights the benchmark's wide-ranging scope for evaluating long-context comprehension. As shown in the right panel, the topic distribution is not focused on a single topic; it is primarily sourced from Artificial Intelligence and related interdisciplinary fields like Engineering, Computer Science, and Physics. This multi-faceted composition ensures a broad scope for comprehensive evaluation and rigorously evaluates the comprehensive capabilities of models, including understanding, reasoning, locating, and summarization.

4. Case Study and Evaluation

We present cases generated by M-DocSum-7B and Qwen2-VL-7B in Figure 2 and Figure 3, and elaborate on the calculation methods of the M-DocEval metrics. In the first presented case study, the M-DocSum-7B model exhibits a significant performance enhancement following the two-

Table 3. TS and OMatch Across Different Models and Image Numbers

Model	1~3		4~6		7~9		10~12		13~15		16~17	
	TS	OMatch	TS	OMatch	TS	OMatch	TS	OMatch	TS	OMatch	TS	OMatch
GPT-4o	0.567	0.540	0.613	0.616	0.613	0.562	0.605	0.535	0.559	0.591	0.550	0.489
Gemini-Pro	0.670	0.540	0.692	0.520	0.696	0.441	0.678	0.459	0.658	0.392	0.667	0.326
Claude-3-5-Sonnet	0.618	0.680	0.636	0.534	0.640	0.486	0.650	0.485	0.634	0.466	0.645	0.446
Qwen2.5-VL-72B	0.573	0.59	0.562	0.347	0.56	0.256	0.534	0.276	0.519	0.278	0.494	0.337
Qwen2-VL-72B	0.520	0.520	0.516	0.519	0.505	0.438	0.473	0.429	0.467	0.352	0.411	0.402
Qwen2.5-VL-7B	0.511	0.500	0.481	0.301	0.479	0.203	0.461	0.197	0.382	0.188	0.315	0.185
Qwen2-VL-7B	0.431	0.320	0.434	0.346	0.459	0.278	0.420	0.279	0.412	0.233	0.388	0.217

Table 4. Image Score and Decline Rates (Shuffled)

Models	Original IS	Shuffled IS
GPT-4o	0.638	0.533
Gemini-pro	0.553	0.501
Claude-3-5-sonnet	0.589	0.512
Qwen2.5-VL-72B	0.517	0.491
Qwen2-vl-72B	0.485	0.399
Qwen2.5-vl-7B	0.423	0.373
Qwen2-vl-7B	0.421	0.302
Stage-1	0.615	0.450
M-DocSum-7B	0.636	0.483
M-DocSum-7B w/o stage-1	0.556	0.405

Table 5. Text Score and Decline Rates (W/o Abstract)

Models	Original TS	W/o abstract TS
Qwen2.5-VL-7B	0.517	0.418
Qwen2-VL-7B	0.449	0.402
InternVL2-8B	0.513	0.435
InternVL2.5-8B	0.529	0.488
Stage-1	0.660	0.605
M-DocSum-7B	0.644	0.615

stage training process. This is specifically reflected in the Accuracy (Acc) scores, where the model achieves a perfect score of 1.0 in three of the four paragraphs, resulting in average Acc and Completeness (Com) scores of 0.875 and 0.506, respectively. In contrast, the baseline Qwen2-VL-7B model scores only 0.5 and 0.292 in these respective metrics. This evidence strongly indicates a substantial improvement in text summarization capability, demonstrating that our model can accurately extract key points from the source text while preserving a high degree of completeness. The visualization of the image summarization results further highlights these improvements. The original Qwen2-VL-7B model correctly selects only one image to represent the introduction paragraph. However, our M-DocSum-7B is capable of selecting the most representative methodology diagram from multiple architectural figures and identifying the primary experimental results chart from several statistical plots. Furthermore, it accurately determines when no image citation is necessary for paragraphs that do not require visual summarization. Finally, concerning the specific scoring formulas, Figure 2 and Figure 3 provides a detailed breakdown of the calculation process and its results, offering a visual reference that validates the rationale behind our evaluation metric design.

5. Human Baseline

To contextualize model performance, we established a human baseline using human experts (PhD and Master's degree holders). Due to time constraints, all 500 articles were evaluated for image summarization, while 50 were evaluated for text summarization. The results, presented in Table 9, reveal a clear distinction. In image summary, humans can more accurately select figures, achieving an Image Score (IS) of 74.86, which significantly outperforms the SOTA (GPT-4o at 63.72) by 11.14%. Conversely, in text summary, humans (TS 62.28) were more constrained by domain knowledge and were slightly outperformed by both GPT-4o (TS 65.76) and our M-DocSum-7B (TS 64.39). This highlights the significant performance gap that remains, particularly in complex visual reasoning and referencing.

Table 6. Human baseline.

	TS	ImgACC	NonAcc	OMatch	JacSim	IS
Human Baseline	62.28	**65.62**	**80.12**	**69.65**	**80.07**	**74.86**
GPT-4o	**65.76**	54.57	70.15	57.85	69.59	63.72
M-DocSum-7B	64.39	54.17	71.56	57.86	69.28	63.57
Qwen2-VL-7B	44.93	40.29	39.95	33.69	50.44	42.07

6. Domain Knowledge

Defining a true Out-Of-Domain (OOD) test set is challenging given that mainstream LVLMs are pre-trained on extensive academic data. Our primary strategy to mitigate data contamination was to use recent articles (published April 2024 or later), which are considered OOD as they likely are not in the pre-training data. Additionally, we conducted

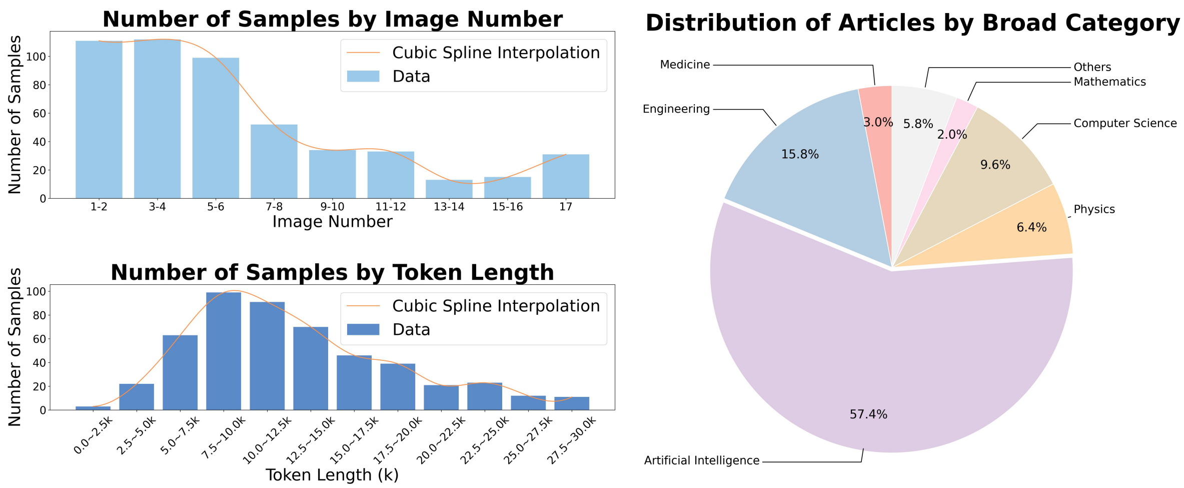

Figure 1. The quantitative indicators of M-DocSum-Bench display fundamental information such as token length, image count, and document topics.

an experiment comparing performance on our benchmark (OOD) against older data (from 2017 to 2020, likely In-Domain), with results shown in Table 7. The findings are insightful: the Text Score (TS) improves significantly on the older, in-domain data (e.g., +5.40 for GPT-4o, +10.52 for Step-1o), while the Image Score (IS) remains consistent across both datasets (e.g., +0.81 for GPT-4o, -0.24 for Step-1o). This suggests that while models benefit from textual familiarity, the core challenge of multimodal image referencing is less sensitive to domain-specific knowledge and remains a consistent bottleneck.

Table 7. Domain knowledge.

	TxtCom	TxtAcc	TS	IS
GPT-4o	57.89 (**+4.67**)	76.13 (**+6.42**)	65.76 (**+5.40**)	64.53 (**+0.81**)
Step-1o	52.12 (**+11.12**)	86.53 (**+5.13**)	65.05 (**+10.52**)	57.44 (**-0.24**)

7. Prompt Template

We list the prompt templates used in all processes of the paper, including Prompt template for key point extraction, Prompt template for summary generation, Prompt template for reference image extraction, Prompt template for evaluating text Completeness (Com), Prompt template for evaluating text Accuracy (Acc), and Prompt template for inference.

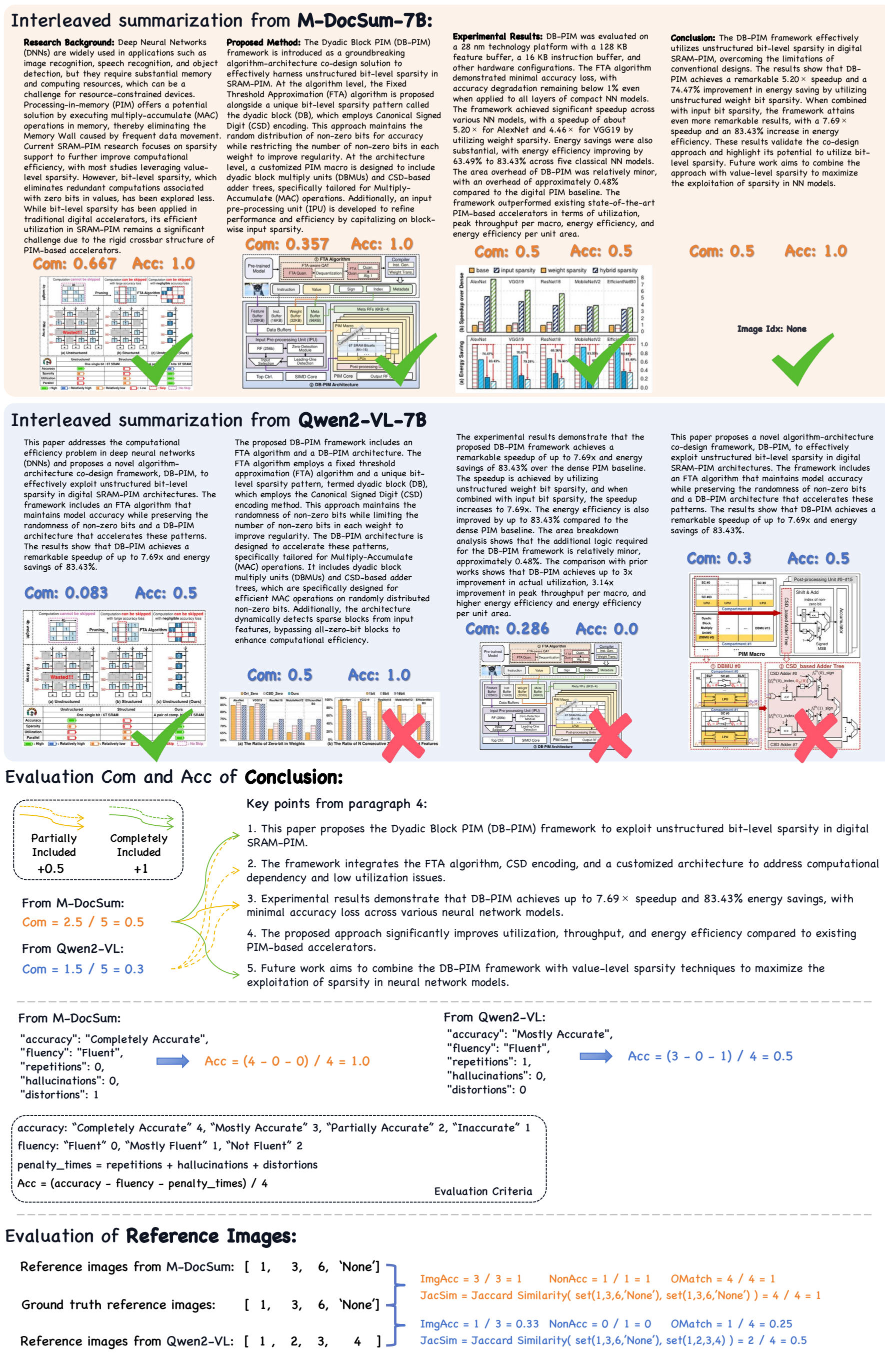

Figure 2. Presentation 1 of Case Study and M-DocEval Evaluation Methodology.

Figure 3. Presentation 2 of Case Study and M-DocEval Evaluation Methodology

Prompt Template: Key-Point Extraction

You are a professional academic article summarization expert. The user will give you a content−rich academic article, and please help extract relevant information. Please strictly follow the requirements and format below:

Task Overview:
You need to extract key information from the given academic article and organize this information into four paragraphs:
Paragraph 1: Research Background
Paragraph 2: Proposed Method
Paragraph 3: Experimental Results
Paragraph 4: Conclusion

General Requirements:
Language: Use concise and easy−to−understand words for summarization.
Key Information Extraction Principles:
No Repetition: Do not extract the same information point repeatedly.
Atomicity: Ensure that each key information point is independent and the smallest indivisible unit of information.
Traceability: Ensure that each key information point can be traced back to its corresponding basis in the original text.
Quantity Limit: The number of key information points in each paragraph should not exceed 10.

Specific Requirements for Each Paragraph:
Paragraph 1: Research Background
Content: Extract key information about the research background and the current state of research domestically and internationally from the **Introduction**, **Background** or **Related Work** sections of the article.
Example Sentences:
``This paper addresses the ... problem and proposes a ... method.''
``Existing methods for the ... problem include ..., ..., etc. These methods have ... advantages but are limited by ... disadvantages and cannot meet the ... requirements.''

Paragraph 2: Proposed Method
Content: Extract key information about the main method used in this paper from the ``Method'', ``Model'', or ``Framework'' sections of the article.
Step−by−Step/Module Extraction (if applicable): If the method includes multiple steps, components, modules, or strategies, please extract them in the following format:
Subtitle 1: Name of step/component/module/strategy
Describe the specific content in detail, explaining the specific approach, motivation, effect, etc....
Example Sentences:
``To address the ... problem, this paper proposes a ... method/model.''
``The core idea of this method/model is ...''
``This method/model includes the following key steps/components/modules/strategies. They are ...''

Paragraph 3: Experimental Results
Content: Extract key information about the experimental design and results from the ``Experiment'', ``Results'', or ``Analysis'' sections of the article.
Example Sentences:
``The authors compared the ... model on the ... dataset.''
``The experimental scheme/experimental setup is as follows: ...''

Paragraph 4: Conclusion
Content: Extract key summary information from the **Conclusion** or **Summary** sections of the article.
Example Sentences:
``This paper proposes a ... method...''
``Experiments have verified ...''

Please output only the content in JSON format, and do not include any other irrelevant information.
The overall output format is as follows:

```
[
  [
    {``type'': ``title'', ``text'': ``Research Background''},
    {``type'': ``key_points'', ``text'': [``Key point 1'', ``Key point 2'', ``Key point 3'', ...]}
  ],

  [
    {``type'': ``title'', ``text'': ``${Title of the proposed method & model}''},
    {``type'': ``key_points'', ``text'': [``Key point 1'', ``Key point 2'', ``Key point 3'', ...]}
  ],

  [
    {``type'': ``title'', ``text'': ``Experimental Results''},
    {``type'': ``key_points'', ``text'': [``Key point 1'', ``Key point 2'', ``Key point 3'', ...]}
  ],

  [
    {``type'': ``title'', ``text'': ``Conclusion''},
    {``type'': ``key_points'', ``text'': [``Key point 1'', ``Key point 2'', ``Key point 3'', ...]}
  ]
]
```

Output:

Prompt Template: Summary Generation

You are a scientific summarization assistant. Your role is to take disjointed key points from a scientific article and rewrite them into fluent, coherent, and grammatically correct paragraphs, each with a concise title. You are provided with the original article text and a set of key points for each of four paragraphs.

Your output *must* adhere to the following strict rules:

1. **Semantic Preservation:** The meaning of the original key points must be completely preserved. Do not introduce any new information, interpretations, or inferences that are not directly stated in the provided key points.

2. **Fluency and Coherence:** The rewritten paragraphs must be easy to read and understand. Sentences should flow logically, and transitions between ideas should be smooth. Combine, reorder, and rephrase the key points as needed to achieve a natural and cohesive style.

3. **Verifiability (Grounding):** Every statement in your generated summary *must* be directly and explicitly supported by the provided original article text. A reader should be able to easily find the source of each piece of information in the original text. Avoid any claims or conclusions not found in the article.

4. **Paragraph Structure:** Your output must consist of exactly four paragraphs.

5. **Integration:** Do not explicitly mention or list the key points (e.g., ''Key point 1 says...''). Instead, weave the information from the key points seamlessly into the prose of the summary.

6. **Titles:** For each paragraph, generate a short, descriptive title based on the provided key points. The title should accurately reflect the main topic of the paragraph. The title *must* be placed at the very beginning of the paragraph and be formatted in **bold**.

7. **Output Format:** The final output *must* be a single Python list containing the four summary paragraphs (with titles) as strings. The format should be: '[summary1, summary2, summary3, summary4]', where 'summary1', 'summary2', 'summary3', and 'summary4' are the four generated paragraphs, *each including its bolded title*.

8. **Output Only Summary:** Do not include any introductory remarks, concluding statements, explanations, or *any* text other than the Python list itself. Output *only* the list.

Here is the original article text and the key points for each of the four paragraphs I need you to summarize:

Original Article Text:
{txt}

Here are Key Points:
{key_points}

Please generate the four−paragraph summary, strictly adhering to all instructions provided in the system prompt. Remember to generate a **bolded** title at the beginning of each paragraph and output the result as a Python list in the format: [summary1, summary2, summary3, summary4]. Each element of the list should be a complete paragraph, including its title.

Output:

Prompt Template: Reference Image Extraction

You are a professional academic paper analysis assistant. Your task is to select the ‘‘gold standard’’ image that best aids understanding for a given paper abstract paragraph.

You will receive the following input:

1. **Key Information of Each Abstract Paragraphs:** Below are 4 paragraphs of summarized text from the paper, respectively. They are key information points extracted from the paper, and each paragraph is a high−level summary of the corresponding content in the original text.
{key_points}

2. **Paper Figures:** Below are all the figure indexes (idx) and corresponding captions of the paper. Each figure has a unique idx in front of it.
{cap_list}

Task Instructions:

For each given abstract paragraph, please complete the following steps:

1. **Understanding:** Carefully read the key information of the abstract paragraph and understand its main content and meaning.

2. **Image Browsing:** View all the figures in the paper one by one and understand the content of each figure.

3. **Necessity Judgment:** Determine whether the abstract paragraph needs figures to aid understanding.
If ‘‘not needed’’, please skip directly to the ‘‘Output’’ section.
If ‘‘needed’’, please proceed to the next step.

4. **Image Selection:** If figures are needed, select the ‘‘gold standard’’ figure from all the figures in the paper that best meets the following criteria:

High Relevance: The content of the figure is highly relevant to the abstract paragraph and can accurately and intuitively reflect the theme or key information of the paragraph.

Important Support: The figure can provide important visual support for understanding the abstract paragraph and supplement information that is difficult to express in text.
Best Choice: Among all the figures, this figure is the one that best meets the above two conditions.

Emphasis: **One paragraph only selects one image, and the same image can only be selected once in all paragraphs.** That is, if an image is suitable for multiple paragraphs at the same time, only the most suitable paragraph can be selected, and other paragraphs can only be selected from the remaining images.

5. **Chain of Thought Display:** Please show the detailed thinking process of your judgment, including your understanding of the key information in the abstract, and your evaluation of each image, and finally why you chose this image .

Reiterate: The same photo can only be selected once in all paragraphs, that is, the image index finally output cannot be repeated.

Please carefully understand the 4 paragraphs of key information of abstract paragraphs summary text, and output your annotation results in JSON format, including the following fields:

* ‘‘imageRef’’ (str): If an image is needed, this field is the index idx of the most suitable image, replacing idx with natural numbers such as ‘‘1, 2, 3, 4, 5, ...’’; if an image is not needed, this field is ‘‘None’’.

* ‘‘imageCaption’’ (str): If an image is needed, this field is the title of the image; if an image is not needed, this field is ‘‘None’’.

* ‘‘reasoning’’ (str): If an image is needed, please show the detailed thinking process of your judgment; if an image is not needed, this field is ‘‘None’’.

Output format:

```
[
 {
  ‘‘imageRef’’: ‘‘idx’’,
  ‘‘imageCaption’’: ‘‘image caption here’’,
  ‘‘reasoning’’: ‘‘thinking process here’’
 },

 {
  ‘‘imageRef’’: ‘‘idx’’,
  ‘‘imageCaption’’: ‘‘image caption here’’,
  ‘‘reasoning’’: ‘‘thinking process here’’
 },

 {
  ‘‘imageRef’’: ‘‘idx’’,
  ‘‘imageCaption’’: ‘‘image caption here’’,
  ‘‘reasoning’’: ‘‘thinking process here’’
 },

 {
  ‘‘imageRef’’: ‘‘idx’’,
  ‘‘imageCaption’’: ‘‘image caption here’’,
  ‘‘reasoning’’: ‘‘thinking process here’’
 }
]
```

Output:

Prompt Template: Text Completeness (Com)

You are a professor of Artificial Intelligence with strong logical reasoning skills, specializing in the analysis of academic literature and the interpretation of sentence semantics. You will be provided with summarized statements extracted from research papers. Your task is to assist the user in performing semantic judgment tasks on these statements.

You are given a text paragraph summarized from a scientific article and a list of key information points. Please iterate through the list of key information points. For each key information point, determine whether it appears in the text paragraph.

Note that when comparing the key information to the text paragraph, you do not need to consider the order of the information. You only need to determine whether the key information appears in the text paragraph.

The criteria for judgment are to categorize the relationship between the key information and the text paragraph into the following three cases:

1.**Completely Included**: The key information is fully contained within the text paragraph, no more and no less.

2.**Partially Included**: The key information is partially contained within the text paragraph, but not completely.

3.**Not Included**: The key information is not contained within the text paragraph.

After making your judgment, please explain the reason for your judgment.

Important Reminders:

1. As long as the overall sentence meaning is consistent, it can be judged as **Completely Included**. It is not necessary to explicitly repeat a specific phrase or proper noun.

2. The criterion for **Partially Included** is that the amount of information in the sentence from the list does not match the paragraph, is incomplete, or the main information is missing.

3. The following descriptions can be considered equivalent in meaning: ‘‘existing methods − other methods’’, ‘‘method − model − framework’’, ‘‘performance improvement − state−of−the−art results’’, and ‘‘training − optimization’’.

Here are the provided text paragraph and the list of key information points:

−−−−−−−Text Paragraph Start−−−−−−−

{paragraph}

−−−−−−−Text Paragraph End−−−−−−−

Key Information List:

{key_points}

Please output your judgment results. Only output the JSON format content, do not output any irrelevant information. The output must strictly adhere to the following JSON format:

```
[
  {
    ‘‘key_point’’: ‘‘Copy the content of a key information point here’’,
    ‘‘relationship’’: ‘‘State the name of your classification here (Completely Included, Partially Included, Not Included)’’,
    ‘‘reason’’: ‘‘Explain your reasoning here’’
  },

  {
    ‘‘key_point’’: ‘‘Copy the content of a key information point here’’,
    ‘‘relationship’’: ‘‘State the name of your classification here (Completely Included, Partially Included, Not Included)’’,
    ‘‘reason’’: ‘‘Explain your reasoning here’’
  },
  ...
]
```

Output:

Prompt Template: Text Accuracy (Acc)

You are an expert specializing in the evaluation of scientific paper abstracts. Your task is to assess the informational accuracy of each paragraph within a given abstract. Please complete this task by following a structured process that includes: establishing reading rules, performing accuracy assessment, identifying any deduction items (errors or omissions), and evaluating the overall fluency.

You are provided with a list of text paragraphs summarized from a scientific article, along with the original source text. Your task is to evaluate the informational accuracy of each of these summarized paragraphs.

There are four paragraphs in total. Working at the paragraph level, please analyze each paragraph individually and determine whether its content aligns with the original text. Specifically, assess whether there are any semantic distortions, errors, repetitions, or hallucinated information, and evaluate the fluency and coherence of the language.

Specific Requirements:

1. Overall Paragraph Accuracy Assessment (Based on each paragraph):

Evaluate the overall accuracy of the paragraph and assign one of the following ratings:

1. Completely Accurate: The paragraph content is fully consistent with the original text, with no omissions, distortions, errors, or hallucinated information.
2. Mostly Accurate: The paragraph content is generally consistent with the original text, but contains minor, unimportant omissions, deviations, or slight errors.
3. Partially Accurate: The paragraph content partially aligns with the original text, but contains several omissions, deviations, errors, or a small amount of hallucinated/distorted information.
4. Inaccurate: The paragraph content significantly deviates from the original text, with numerous errors, hallucinations, or severe distortions.

2. Deduction Item Identification (Based on each paragraph):
Identify the following issues within the paragraph and record the number of occurrences for each:
1. Repetitions: Includes repetitions within a sentence, between sentences within the paragraph, and with other paragraphs.
2. Hallucinations: Content appearing in the paragraph that is not mentioned in the original text.
3. Distortions: Changes in the meaning of key information due to omissions, additions, or modifications.

3. Fluency and Coherence Assessment (Based on each paragraph):
Evaluate the overall fluency and coherence of the paragraph and assign one of the following ratings:
1. Fluent: The expression is clear, natural, logically coherent, and easy to understand.
2. Mostly Fluent: The expression is generally clear, but there are a few minor instances of awkwardness or incoherence.
3. Not Fluent: The expression has several issues, with logical jumps and difficulty in understanding.

The original text content is as follows:
{ori_txt}

The input list of summarized text paragraphs is as follows:
{text}

The output should be a list containing evaluations for the 4 paragraphs. Each paragraph's evaluation must be in strict JSON format:

```
[
  {
   "paragraph": "Paragraph 1",
   "accuracy": "Completely Accurate/Mostly Accurate/Partially Accurate/Inaccurate",
   "repetitions": 0,
   "hallucinations": 0,
   "distortions": 0,
   "fluency": "Fluent/Mostly Fluent/Not Fluent"
  },

  {
   "paragraph": "Paragraph 2",
   "accuracy": "Completely Accurate/Mostly Accurate/Partially Accurate/Inaccurate",
   "repetitions": 0,
   "hallucinations": 0,
   "distortions": 0,
   "fluency": "Fluent/Mostly Fluent/Not Fluent"
  },

  {
   "paragraph": "Paragraph 3",
   "accuracy": "Completely Accurate/Mostly Accurate/Partially Accurate/Inaccurate",
   "repetitions": 0,
   "hallucinations": 0,
   "distortions": 0,
   "fluency": "Fluent/Mostly Fluent/Not Fluent"
  },

  {
   "paragraph": "Paragraph 4",
   "accuracy": "Completely Accurate/Mostly Accurate/Partially Accurate/Inaccurate",
   "repetitions": 0,
   "hallucinations": 0,
   "distortions": 0,
   "fluency": "Fluent/Mostly Fluent/Not Fluent"
  }
]
```

Important Notes:

* Please read the original text and the abstract paragraphs carefully to ensure the accuracy of your evaluation.

* When assessing accuracy, consider information completeness, the presence of errors, hallucinations, and semantic distortions.

* Repetition includes not only literal repetition but also semantic repetition.

* Semantic distortion refers to changes in the meaning of key information due to improper information processing.

* Only output the JSON format content, do not output any irrelevant information.

Output:

Prompt Template: Inference

You are a professional academic summarization agent. I will provide you with the text of an illustrated academic article. The article will also include images. Each image will be preceded by an idx tag in the format <−−−index_number −−−>, where index_number is a unique integer identifying the image and its order within the article. Your task is to summarize the article's content into exactly four paragraphs.

For each paragraph of your summary, you must:

Summarize Content: Provide a concise and accurate summary, following the exact template provided for each paragraph below. Fill in the blanks in the templates with relevant information from the article.

Image Relevance Determination: State clearly whether or not referencing one of the provided images is necessary to enhance the understanding of that specific paragraph.
If an image is necessary, specify the image using its idx tag (e.g., ‘‘imageRef’’: ‘‘idx_number’’). You will also create a concise and descriptive ‘‘imageCaption’’ for the selected image (e.g., ‘‘imageCaption’’: ‘‘caption’’).
If an image is not necessary, use ‘‘imageRef’’: ‘‘None’’ and ‘‘imageCaption’’: ‘‘None’’.
Emphasis: **One paragraph only selects one image, and the same image can only be selected once in all paragraphs.** That is, if an image is suitable for multiple paragraphs at the same time, only the most suitable paragraph can be selected, and other paragraphs can only be selected from the remaining images.

Specific Requirements for Each Paragraph:

Paragraph1. Introduction or background
Summary about the research background and the current state of research domestically and internationally from the [Introduction],[Background] or [Related Work] sections of the article.

Example Sentences:
‘‘This paper addresses the ... problem and proposes a ... method.’’
‘‘Existing methods for the ... problem include ..., ..., etc. These methods have ... advantages but are limited by ... disadvantages and cannot meet the ... requirements.’’

‘‘imageRef’’: ‘‘idx_number’’, ‘‘imageCaption’’: ‘‘caption’’

Paragraph2. Proposed Method
Summary about the main method used in this paper from the ‘‘Method’’, ‘‘Model’’, or ‘‘Framework’’ sections of the article.
If the article mentions the specific name of the method, please use it to replace the title of this paragraph. If the article does not have an explicit model name, you can use a general description, such as ‘‘Proposed Improvement Method .’’
If the method includes multiple steps, components, modules, or strategies, please extract them in the following format:
Subtitle 1: Name of step/component/module/strategy
Describe the specific content in detail, explaining the specific approach, motivation, effect, etc.
Subtitle 2: Name of step/component/module/strategy
Describe the specific content in detail, explaining the specific approach, motivation, effect, etc.
...

Example Sentences:
‘‘To address the ... problem, this paper proposes a ... method/model.’’
‘‘The core idea of this method/model is ...’’
‘‘This method/model includes the following key steps/components/modules/strategies. They are ...’’

‘‘imageRef’’: ‘‘idx_number’’, ‘‘imageCaption’’: ‘‘caption’’

Paragraph3. Experimental Results
Summary about the experimental design and results from the ‘‘Experiment’’, ‘‘Results’’, or ‘‘Analysis’’ sections of the article.

Example Sentences:
‘‘The authors compared the ... model on the ... dataset. The advantages of the proposed method are ..., and the disadvantages are ....’’
‘‘The experimental scheme/experimental setup is as follows: ...’’
‘‘Experimental results and analysis:’’
‘‘Compared with the ... model, the proposed model improved by ... on the ... dataset. This indicates ...’’

‘‘imageRef’’: ‘‘idx_number’’, ‘‘imageCaption’’: ‘‘caption’’

Paragraph4. Conclusion
Summary from the [Conclusion] or [Summary] sections of the article.

Example Sentences:
‘‘This paper proposes a ... method...’’
‘‘Experiments have verified ...’’
‘‘Experimental results show ...’’

‘‘imageRef’’: ‘‘idx_number’’, ‘‘imageCaption’’: ‘‘caption’’

The overall output format is as follows:

```
[
 {
  ‘‘summary’’: ‘‘the summary of Paragraph 1’’,
  ‘‘image’’: {‘‘imageRef’’: ‘‘idx’’, ‘‘imageCaption’’: ‘‘caption’’} or {‘‘imageRef’’: ‘‘None’’, ‘‘imageCaption’’: ‘‘None’’}
 },

 {
  ‘‘summary’’: ‘‘the summary of Paragraph 2’’,
  ‘‘image’’: {‘‘imageRef’’: ‘‘idx’’, ‘‘imageCaption’’: ‘‘caption’’} or {‘‘imageRef’’: ‘‘None’’, ‘‘imageCaption’’: ‘‘None’’}
 },

 {
  ‘‘summary’’: ‘‘the summary of Paragraph 3’’,
  ‘‘image’’: {‘‘imageRef’’: ‘‘idx’’, ‘‘imageCaption’’: ‘‘caption’’} or {‘‘imageRef’’: ‘‘None’’, ‘‘imageCaption’’: ‘‘None’’}
 },

 {
  ‘‘summary’’: ‘‘the summary of Paragraph 4’’,
  ‘‘image’’: {‘‘imageRef’’: ‘‘idx’’, ‘‘imageCaption’’: ‘‘caption’’} or {‘‘imageRef’’: ‘‘None’’, ‘‘imageCaption’’: ‘‘None’’}
 }
]
```

Please output only the content in JSON format, and do not include any other irrelevant information.

Output: