Multimodal Knowledge Retrieval-Augmented Iterative Alignment for Satellite Commonsense Conversation

Qian Li[1*], **Xuchen Li**[2], **Zongyu Chang**[1], **Yuzheng Zhang**[1], **Cheng Ji**[3] and **Shangguang Wang**[1]
[1]School of Computer Science, Beijing University of Posts and Telecommunications, China
[2]Institute of Automation, Chinese Academy of Sciences and Zhongguancun Academy, China
[3]SKLCCSE, School of Computer Science and Engineering, Beihang University, China
li.qian@bupt.edu.cn, lixuchen2024@ia.ac.cn, changzongyu@bupt.edu.cn, zhangyuzheng@bupt.edu.cn, jicheng@act.buaa.edu.cn, sgwang@bupt.edu.cn

Abstract

Satellite technology has significantly influenced our daily lives, manifested in applications such as navigation and communication. With its development, a vast amount of multimodal satellite commonsense data has been generated, thus leading to an urgent demand for conversation about satellite data. However, existing large language models suffer from prevalent hallucinations and poor comprehensibility on multimodal satellite data due to their high professional content threshold and partial information opacity. To address these issues, we propose a multimodal satellite knowledge retrieval-augmented iterative alignment framework (Sat-RIA) for satellite commonsense conversation. We first construct multi-view retrieval expert knowledge to reduce hallucinations and enhance the interpretability of responses, which incorporates the satellite expert database, satellite rule, satellite image database, and a satellite knowledge graph. We next design commonsense conversation instructions to make the answers more legible and understandable. Furthermore, the retrieval-augmented iterative alignment module refines response precision by aligning outputs with task-specific standards through multi-stage evaluations. Finally, we construct satellite multi-turn dialogue and visual question-answer datasets for a more comprehensive evaluation of satellite commonsense conversation. Experimental results demonstrate that Sat-RIA outperforms existing large language models and provides more comprehensible answers with fewer hallucinations.

1 Introduction

Satellite technology is integral to modern telecommunications, earth observation, and global navigation, with its applications spanning critical areas such as climate change monitoring, disaster management, and the security of global communications networks [Chen *et al.*, 2023a; Lu *et al.*, 2021]. The vast amounts of data generated by satellites offer unprecedented insights into environmental patterns, human activities, and natural phenomena, making them indispensable tools for scientific research, government operations, and commercial ventures [Choi, 2024; Inkollu and Sastry, 2024].

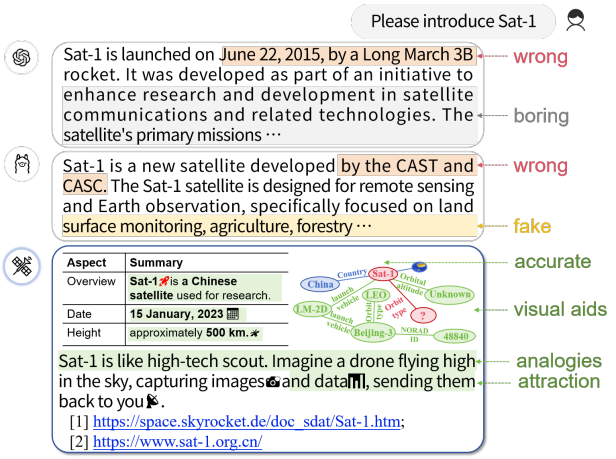


Figure 1: Examples of GPT-4o, LLaMa3, and Sat-RIA (Ours). For satellite knowledge, GPT-4o and LLaMa3 easily generate fake results, which are marked in red color. Compared to others, our Sat-RIA generates more accurate and comprehensible results with fewer hallucinations and provides abundant ways to read easily.

With the increasing sophistication of satellite technologies, there is a growing demand for advanced digital platforms and web-based tools capable of efficiently disseminating satellite knowledge to a wider, more diverse audience, including non-experts [Gallardo, 2024; Li *et al.*, 2024a]. Traditional methods of sharing satellite data, such as technical reports and static datasets, often fail to engage non-specialists and limit the democratization of satellite knowledge. Despite the potential of large language models (LLMs) [Chang *et al.*, 2024; Kasneci *et al.*, 2023] to simplify and popularize satellite knowledge, existing LLM models [Thirunavukarasu *et al.*, 2023] face significant limitations when applied to domain-specific data, as shown in Figure 1. Firstly, LLMs struggle with hallucinations, especially in specialized fields like satellite technology, due to a lack of satellite-specific knowledge in their pre-training and tuning phases. This leads to the generation of false or irrelevant information, undermining their reliability. Secondly, the outputs from current models often lack engagement, partic-

∗ Corresponding author.

ularly for non-expert users [Segovia-Dominguez *et al.*, 2021; Gadiraju *et al.*, 2020], as they tend to be monotonous, redundant, and devoid of interactivity.

To avoid the erroneous dissemination of satellite common knowledge by existing LLMs and to enhance the understandability and appeal of the generated results, we propose Sat-RIA, a multimodal satellite knowledge retrieval-augmented iterative alignment framework for satellite commonsense conversation. including the multi-turn dialogue and visual question-answer tasks. Specifically, we design multi-view retrieval expert knowledge construction that integrates our constructed satellite knowledge graph to reduce token consumption and combines the satellite expert database, satellite rule, and a satellite image database, which allows the model to retrieve and apply domain-specific information during both training and inference. We also design commonsense conversation instructions to make the answers more legible and understandable. To further ensure the reliability of the model's outputs, we design a retrieval-augmented iterative alignment module, that employs a multi-stage evaluation process that continually refines the model's performance by aligning its generated responses with established task-specific standards. This iterative refinement reduces errors and enhances the model's ability to handle complex satellite tasks with greater precision. Finally, we construct satellite commonsense multi-turn dialogue and visual question-answer datasets for satellite commonsense conversation evaluation. Experimental results demonstrate that our approach outperforms existing large language models in answering satellite-commonsense-related questions and provides more comprehensible answers with fewer hallucinations. Our key contributions can be summarized as follows:

- To our best knowledge, we are the first to propose a novel multimodal knowledge retrieval-augmented iterative alignment framework for satellite commonsense conversation to provide more comprehensible answers with fewer hallucinations generation results.
- We design a multi-view retrieval expert knowledge construction to reduce model hallucinations and enhance the interpretability of responses, alongside a retrieval-augmented iterative alignment mechanism that refines outputs through multi-stage evaluations, ensuring task-specific reliability.
- We construct a satellite knowledge graph, satellite expert database and a multi-modal dataset containing satellite multi-turn dialogue and visual question-answer for training and evaluation across both textual and visual tasks.

2 Related Work

Research on applying language models to satellite technology has primarily focused on data classification and anomaly detection [Bondi *et al.*, 2022; Chen *et al.*, 2021; de Witt *et al.*, 2021; Li *et al.*, 2023]. However, these applications have not fully exploited the potential of integrating these models with knowledge graphs to enhance their reasoning capabilities [Han *et al.*, 2020; Lebedev *et al.*, 2019]. While large language models [Zhou *et al.*, 2023; Kalyan, 2024; Huang and Chang, 2023] have shown promise in understanding satellite data, there exists a significant issue of hallucination and a lack of interpretability in their understanding of satellite common knowledge. In addition, we supplement the related work on Multi-Modal LLMs and Domain Large Language Models in the Appendix A.

3 Preliminaries

Satellite commonsense conversation consists of two different tasks – satellite commonsense multi-turn dialogue (SatDiag) and satellite commonsense visual question-answer (SatVQA).

Definition 1 (Satellite Commonsense Multi-Turn Dialogue (SatDiag)). *Given a textual query q represented as a set of words, the objective of the SatDiag task is to generate a response r that is coherent and relevant to the context c. The task can be defined as a function $f_{SatDiag} : (q, c) \rightarrow r$, where $q \in \mathcal{Q}$ is a textual query from the set of all possible queries $\mathcal{Q}$, $c \in \mathcal{C}$ is the context from the set of all possible contexts $\mathcal{C}$, and $r \in \mathcal{R}$ is the generated response from the set of all possible responses $\mathcal{R}$.*

Definition 2 (Satellite Commonsense Visual Question-Answer (SatVQA)). *Given a textual query q and a satellite image i, the task is to generate a detailed textual description to answer the question q regarding the satellite image i. It is worth noting that satellite images here refer to pictures of the satellite itself rather than images taken by the satellite. The task can be defined as a function $f_{SatVQA} : (q, i) \rightarrow d$, where $q \in \mathcal{Q}$ is a textual query from the set of all possible queries $\mathcal{Q}$, $i \in \mathcal{I}$ is a satellite image from the set of all possible satellite images $\mathcal{I}$, and $d \in \mathcal{D}$ is the generated description from the set of all possible descriptions $\mathcal{D}$.*

4 Sat-RIA Framework

LLMs are prone to hallucination, especially in domain-specific fields where professional terminology may be confused or misinterpreted. To address this, we propose a satellite commonsense multimodal knowledge retrieval-augmented iterative alignment framework (Sat-RIA) that not only inputs the query q but also retrieves relevant information from the satellite database s, related parameter satellite design rules r, satellite images i, and the knowledge graph $\mathcal{G}$, as shown in Figure 2. This retrieval-augmented iterative alignment mechanism reduces errors and enhances the model's ability to handle complex satellite tasks with greater precision into the Sat-RIA, enhancing response accuracy and minimizing hallucinations.

4.1 Multi-View Retrieval Knowledge Construction

To effectively support the satellite-specific tasks of Sat-RIA, we propose the multi-view retrieval knowledge construction. It integrates data from various sources, including satellite databases, expert knowledge, and visual datasets, to ensure a robust and contextually rich knowledge base. The multi-view retrieval knowledge construction enables the model to dynamically retrieve relevant information across multiple modalities, such as text, images, and structured satellite data, ensuring accurate and informed outputs.

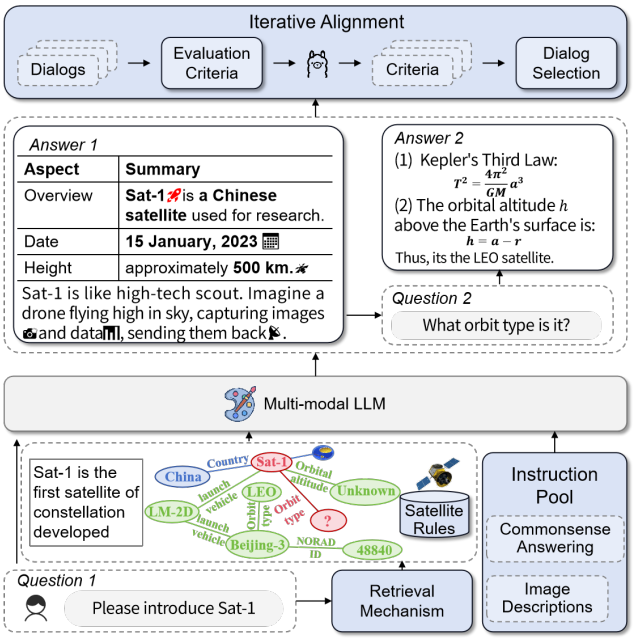


Figure 2: The framework of the Sat-RIA.

Satellite Expert Database Construction. To address the limitations of existing satellite datasets, we propose the construction of an extensive satellite expert database. Initially, data were collected from Nanosats.eu, which provides information on 2,064 satellites. However, recognizing the limitations of this dataset, we expanded the database by utilizing Google's API to query over 7,000 satellites. The data were further enriched with key parameters from curated expert satellite sources to ensure comprehensive data quality. To resolve discrepancies between multiple data sources, we employ a weighted voting mechanism, defined as follows:

$$P^* = \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} \cdot v_i, \tag{1}$$

where P^* represents the final parameter value, v_i denotes the value from the i-th source, and w_i corresponds to the weight assigned to that source based on its reliability. This mechanism generates a high-quality, comprehensive satellite database for LLM usage.

Satellite Rule Construction. We propose the design of a Satellite Rule Collection to integrate domain-specific expert knowledge and ensure accurate reasoning in satellite applications. This collection includes a Satellite Design Rules Library, containing 41 guidelines on structural integrity, thermal control, power management, and communication systems, alongside a Satellite Design Formulas Library, comprising 48 key formulas related to orbital mechanics and thermal dynamics (e.g., Kepler's equations). To ensure the robustness and quality of these rules, a validation mechanism was developed, integrating both physical and engineering constraints. The consistency-checking algorithm cross-validates each rule within the Satellite Design Rules Library against these constraints, with the validation process governed by the following scoring function:

$$S_{\text{rule}}\ (r_i) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{F}\left(r_i, c_k\right), \tag{2}$$

where r_i denotes the satellite design rule under evaluation, and c_k represents the k-th physical or engineering constraint. The function $\mathcal{F}\left(r_i, c_k\right)$ outputs a binary value indicating whether the rule r_i satisfies constraint c_k. A score closer to 1 signifies higher compliance with essential constraints. Additionally, these rules are verified against real-world satellite data, enhancing both their reliability and applicability.

Satellite Knowledge Graph Construction To systematically organize satellite data, we propose the construction of a comprehensive Satellite Knowledge Graph (more details in Appendix B). This knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ integrates structured data, such as design parameters, mission objectives, operational status, and historical records. The graph comprises nodes $\mathcal{V}$ (e.g., satellite, launch vehicle) and edges $\mathcal{E}$ (e.g., *launched by*, *operated by*). To ensure the integrity and accuracy of the knowledge graph, we propose a consistency-check mechanism that validates the relationships and data within the graph using multi-relational embeddings. The relationships between entities v_k via relationship e_j are validated with the following scoring function:

$$f\left(v_i, e_j, v_k\right) = \left|\mathbf{v}_i + \mathbf{e}_j - \mathbf{v}_k\right|^2. \tag{3}$$

It ensures that relationships are consistent and represent valid connections between satellite components, mission data, and operational parameters. Relationships exceeding a threshold of k are flagged for further review.

Satellite Image Database Construction. We propose the construction of a comprehensive satellite image dataset, encompassing visual data on satellites and launch vehicles. The dataset is compiled through two primary sources: the Google Images Crawler, which retrieves the three most relevant images for each entity using image recognition algorithms, and the Nanosats.eu Crawler, which extracts high-quality images from the Nanosats.eu website, including detailed specifications and mission descriptions. To ensure the relevance and integrity of the image data, we introduce a cross-validation algorithm that evaluates image quality using the Structural Similarity Index (SSI) [Wang *et al.*, 2003] between retrieved images and reference images from trusted databases:

$$\text{SSI} = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{4}$$

where μ_x and μ_y are the means of the reference and retrieved images, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance between the images. By maximizing the SSI, we ensure that the retrieved satellite images are of high quality and relevant for downstream tasks, such as satellite image analysis.

4.2 Commonsense Conversation Instruction

To ensure comprehensive and robust handling of satellite-related tasks, we design and implement a multi-faceted approach that integrates both dialogue and visual question-answer tasks. It allows Sat-RIA to effectively process and generate high-quality responses, tailored to both experts and non-experts in the satellite domain.

SatDiag Instruction. We design a prompt-based satellite multi-turn dialogue system that facilitates natural and coherent interactions between users and Sat-RIA. The key to effective dialogue generation lies in the precise design of prompts, ensuring each interaction targets specific satellite-related knowledge. To elicit accurate satellite-related information, we design prompts for question-answering tasks that guide the model towards retrieving specific knowledge from the satellite knowledge graph. This enables the model to generate factually accurate and contextually relevant responses. A typical prompt for question answering would be:

> *"Given the specific context provided by the retrieval knowledge [knowledge], how would you address the following query: [query]?"*

We next design prompts for multi-turn dialogues to simulate realistic, ongoing conversations about satellite technology. These dialogues allow the model to maintain a coherent flow while addressing multiple aspects of satellite operations or research in a single conversation. A sample prompt for this type of dialogue might be:

> *"Continuing from the previous discussion on satellite technology, how would you address the following point: [next topic]?"*

This design ensures that the model can engage in up to 10 exchanges per satellite, offering a dynamic and engaging dialogue experience. To facilitate the accessibility and comprehension of satellite-related concepts for a diverse audience, we propose a series of commonsense conversation instruction. These instructions focus on enhancing the clarity and engagement of model-generated responses, ensuring that complex ideas are communicated effectively.

1. Tabular Format: *Present responses in tabular form to improve clarity and comprehension.*
2. Simplified Language: *Use simple language and avoid technical jargon to ensure that technical terms are easily understood.*
3. Use of Analogies: *Incorporate everyday analogies to simplify complex satellite concepts.*
4. Storytelling: *Engage users with narratives, either real or fictional, to illustrate satellite technologies or discoveries.*
5. Proof Links: *Provide references to authoritative literature or official sources to enhance verification and credibility.*
6. Highlighting Key Points: *Highlight critical information with bold text or larger fonts, and incorporate engaging elements like emoticons.*
7. Fine-Grained Perception: *Identify detailed characteristics, such as color, material, and shape, for a more comprehensive understanding.*

By structuring responses in this way, Sat-RIA is designed to effectively communicate complex satellite knowledge, making it more accessible and engaging for a broad audience.

SatVQA Instruction. We propose a detailed multi-step process for constructing satellite visual question-answer to ensure diversity, accuracy, and depth in image interpretation. The construction of this dataset involves collecting a diverse range of satellite images and using advanced language models to generate comprehensive descriptions. In the first step, we use GPT-4V to generate detailed descriptions for each satellite image. These descriptions are crafted using a variety of prompts that focus on different visual and contextual aspects of the image. For example, a prompt might ask the model to describe the visual experience evoked by the image, considering factors such as style, theme, and setting:

> *"Please observe and describe the experience or feelings elicited by this picture, discussing aspects such as style, theme, setting, mood, and quality."*

We further propose a draft creation process wherein the generated descriptions are consolidated with the original descriptions provided on satellite-related websites. This is achieved by instructing GPT-4V to merge information from multiple sources, ensuring consistency and accuracy across descriptions. The prompt used for this task is as follows:

> *"You are a text information integration expert. Merge the information from two texts describing an image from different perspectives into one comprehensive and detailed description. Retain as much valid information from both texts as possible. Ensure the integrated text is accurate and consistent with the original descriptions."*

To ensure accuracy and comprehensiveness, the initial drafts are reviewed by multiple human annotators. These experts refine the descriptions by adding additional details and correcting any errors in multiple rounds of verification, ensuring that the final text is both precise and exhaustive. The final dataset includes 1-5 description dialogues for each satellite image, ensuring depth and variability. Finally, we propose a structured methodology for generating detailed visual question-answers. These descriptions aim to provide enough detail to allow individuals unfamiliar with the image to accurately reconstruct its content based solely on the description.

1. Accuracy: Ensure that descriptions are detailed, factually correct, and free from misinformation. Provide specifics such as object names (e.g., satellite names) and their attributes (e.g., color, material, shape), and summarize key points in a table for clarity.
2. Question Splitting: For complex queries, divide the description into sub-questions, ensuring that each is addressed thoroughly.
3. Proof Links: Include links to authoritative scientific sources and official websites to ensure the credibility of the description.
4. Highlighting Key Points: Use bold text or larger fonts to emphasize key information, and incorporate engaging elements like emoticons where appropriate.
5. Length: Ensure each visual question-answer is at least 300 words in length, providing sufficient detail for a comprehensive understanding.

4.3 Retrieval-Augmented Iterative Alignment

Retrieval. To optimize response accuracy, we propose a retrieval mechanism that dynamically fetches relevant information from the satellite knowledge graph. The relevance between the query Q and the knowledge graph nodes K is computed using the following function:

$$\mathrm{Re}(Q, K) = \mathrm{Sim}(Q, K) + \mathrm{W}(K, Q) \cdot \mathrm{Imp}(K, \mathcal{G}), \quad (5)$$

where Q is the query vector, K represents nodes in the knowledge graph, and $\mathrm{Sim}(Q, K)$ calculates the similarity between the query and the nodes using the cosine similarity function. $\mathrm{W}(K, Q)$ is a weighting function that quantifies node relevance to the query, and $\mathrm{Imp}(K, \mathcal{G})$ reflects the node's importance within the graph based on its connectivity and strategic position. During inference, these scores are used to retrieve the most pertinent information, ensuring contextually aligned and well-informed responses.

Iterative Alignment. We propose a deconfounded strategy to generate alignment responses during training, allowing the model to progressively align its outputs with the expected form. The model generates multiple candidate responses $\{y_1, y_2, \ldots, y_n\}$ using sampling with different random seeds, while keeping the input x and decoding parameters constant. We also propose an evaluation mechanism based on large locally deployed models (e.g., LLaMa), which use five evaluation criteria to assess each candidate's response. The best response is selected based on a comprehensive score, thereby allowing the model to produce more refined outputs during training. To reduce the cost of data collection at each optimization step, we propose an iterative alignment approach. At each iteration, N multi-modal instructions are selected, and the deconfounded strategy is used to generate n candidate responses with the current model M_i. Each response is evaluated using a locally deployed model L, and the resulting alignment data D_i are used to further optimize the model, resulting in M_{i+1}, which forms the basis for the next iteration. This iterative process continuously improves the model's performance, particularly for multi-modal tasks, while minimizing data collection costs. We implement a weighted optimization approach for alignment, minimizing the discrepancy between the model's output and the expected output with the following formula:

$$M_{i+1} = M_i - \eta \cdot \frac{1}{n} \sum_{j=1}^{n} w_j \cdot \nabla_\theta L\left(y_j, x, M_i\right), \quad (6)$$

where M_i is the model at iteration i, η is the learning rate, y_j is the j-th candidate response, w_j is the weight assigned to y_j based on its comprehensive score $S(y_j)$, and $L\left(y_j, x, M_i\right)$ is the loss function that measures the discrepancy between y_j and the expected output. The comprehensive score $S(y_j)$ for each response is computed as:

$$S\left(y_j\right) = \frac{1}{m} \sum_{k=1}^{m} c_k\left(y_j\right), \quad (7)$$

where $c_k(y_j)$ represents score for the k-th evaluation criterion, which includes m metrics such as consistency, correctness, context relevance, detail orientation, and keyword accuracy.

5 Experiments

5.1 Evaluation Datasets

To evaluate our models on satellite commonsense conversation, we construct two datasets: one for satellite multi-turn dialogues (SatDiag) and one for satellite visual question-answering (SatVQA) (more details in Appendix C). The SatDiag dataset includes 2,000 dialogues focused on satellite operations and related knowledge, covering topics like satellite design, formulas, and parameters. Each dialogue is annotated with intent, key entities, and contextual information to assess contextual understanding and response accuracy. The SatVQA dataset consists of 2,000 labeled examples describing satellite images from various scenarios, including components, configurations, and operations. Annotations highlight the primary subjects, notable features, and context, enabling evaluation of multi-modal understanding.

5.2 Evaluation Metrics

To evaluate model performance on SatVQA and SatDiag datasets, we use both traditional and custom metrics (more details in Appendix D). Traditional metrics include BLEU, METEOR, GLEU, and CIDEr, which assess the quality, coherence, relevance, and accuracy of generated text. Custom metrics, evaluated by ChatGPT-4v, measure four dimensions: 1) **Consistency:** Ensures the response is free from contradictions. 2) **Context:** Assesses how well the response fits the broader context. 3) **Correctness:** Evaluates the factual accuracy of the response. 4) **Detail:** Measures the richness and specificity of the details provided. Each criterion is scored from 0 to 10, with higher scores indicating better performance. Additionally, we evaluate keyword hits using recall, precision, and F1 scores to assess the model's ability to incorporate relevant satellite-related terms.

5.3 Comparison Methods

We compare our model with several state-of-the-art LLM baselines (more details in Appendix E): 1) **InternVL 2** [Chen *et al.*, 2023b] is a multi-modal model with 8B parameters, excelling in visual question-answer tasks but slightly weaker in dialogue generation. 2) **LLaVa 1.6** [Liu *et al.*, 2023b; Liu *et al.*, 2023a] is a 7B-parameter model focused on fine-grained vision-language alignment, performing well in visual question-answer. 3) **Deepseek-VL** [Lu *et al.*, 2024] uses 7B parameters for complex image and dialogue tasks, excelling in detail-rich scenes. 4) **Yi-VL** [Young *et al.*, 2024] is a 6B-parameter model optimized for fast dialogue generation with efficient handling of vision language tasks.

5.4 Implementation Details

The Sat-RIA was trained using the PyTorch framework, leveraging the computational power of NVIDIA GPUs to expedite the training process. we use InternVL 2 8B [Chen *et al.*, 2023b] as the foundational LLM provided a strong baseline for our fine-tuning efforts. We use a total batch size of 1 throughout the training process. The AdamW [Loshchilov and Hutter, 2019] optimizer is applied with a cosine learning rate decay and a warm-up period. In the training stage, every alignment epoch number is 1 with a learning rate of 1×10^{-5}

Model	Size	Satellite Multi-Turn Dialogue (SatDiag)				Satellite Visual Question-Answer (SatVQA)			
		BLEU	METEOR	GLEU	CIDEr	BLEU	METEOR	GLEU	CIDEr
LLaVA 1.6	7B	4.71	24.41	7.17	5.35	1.29	20.78	4.67	6.92
Deepseek-VL	7B	8.20	21.41	11.33	6.42	2.27	15.17	6.92	6.85
Yi-VL	6B	3.10	14.07	8.46	4.97	3.37	18.00	7.87	<u>7.42</u>
InternVL 2	8B	<u>8.27</u>	<u>26.28</u>	<u>11.38</u>	<u>9.92</u>	<u>5.92</u>	<u>24.06</u>	<u>10.41</u>	7.41
Sat-RIA (Ours)	8B	**14.86 (↑6.59)**	**36.18 (↑9.90)**	**17.74 (↑6.36)**	**11.75 (↑1.83)**	**17.64 (↑11.72)**	**44.41 (↑20.35)**	**21.60 (↑11.19)**	**10.54 (↑3.12)**

Table 1: Performance comparison of different baseline models. The best results are highlighted in bold and the underlined values are the second-best result. "↑" means the increase compared to the underlined values.

Variant	Satellite Multi-Turn Dialogue (SatDiag)				Satellite Visual Question-Answer (SatVQA)			
	Consistency	Context	Correctness	Detail	Consistency	Context	Correctness	Detail
Sat-RIA	**9.78**	**9.78**	**9.83**	**8.17**	**7.51**	**8.30**	**7.58**	**6.61**
w/o Satellite Rule	9.39	9.56	9.78	7.65	7.29	8.09	7.51	6.16
w/o Satellite Expert Database	3.00	2.78	3.09	2.22	4.72	5.20	6.26	5.18
w/o Iterative Alignment	9.48	9.61	9.43	7.65	7.39	8.22	7.49	6.47
w/o Retrieval Mechanism	2.04	3.09	2.22	2.78	4.97	5.03	6.54	5.07

Table 2: Variant experiments evaluating by ChatGPT-4v. "w/o" means removing the corresponding module from complete model.

and a warmup ratio of 0.05. Hyperparameters were fine-tuned iteratively based on the performance metrics observed during validation. We have trained our model through the method of full parameter fine-tuning, using a 2xA800 80G machine, and All experiments were conducted on the same machine. For alignment response evaluation, we use the locally deployed LLaMa3 8B model (more details in Appendix F).

5.5 Main Results

To verify the effectiveness of our model, we report the overall average results in Table 1. We experimented with the traditional metrics. From the table, we can observe that: 1) Our model outperforms all others across metrics in both the Satellite Multi-turn Dialogue and Satellite Visual Question-Answer tasks, indicating its superiority in generating accurate and relevant satellite content. 2) In Satellite Multi-turn Dialogue, our model achieves the highest BLEU score of 14.86, well above InternVL 2's 8.27, demonstrating its ability to generate fluent dialogues through satellite expert data and retrieval. 3) Our model's METEOR score of 36.18 for Satellite Multi-turn Dialogue reflects its strength in generating semantically meaningful responses. 4) For Satellite Visual Question-Answer, our model leads with a BLEU score of 17.64, showcasing its capability to produce detailed descriptions. 5) It also excels in METEOR, GLEU, and CIDEr metrics, with scores of 44.41, 21.60, and 10.54, respectively, underlines the robustness and effectiveness of our approach in handling diverse satellite-related information generation tasks. 6) Compared to baseline models like InternVL 2 and LLaVa 1.6, our model consistently achieves higher scores, confirming the effectiveness of our proposed architecture. All the observations demonstrate the effectiveness of the Sat-RIA framework.

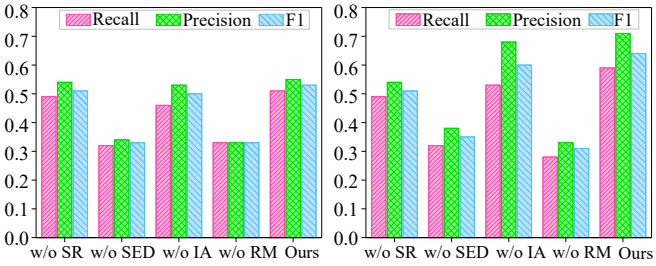
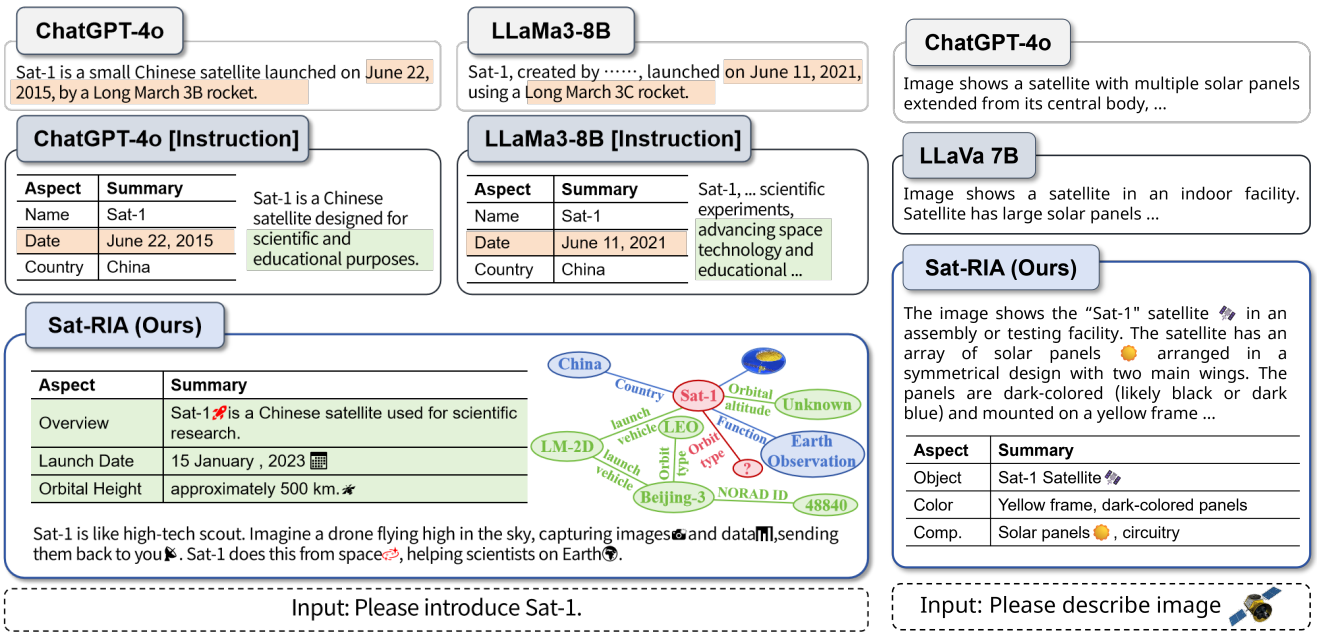


Figure 3: Variant experiments on evaluating keyword hits.

5.6 Discussion for Model Variants

To investigate the effectiveness of each module in Sat-RIA, we conducted a series of variant experiments, with the results presented in Tables 2 and Figure 3. We evaluated performance using two different types of metrics. In Table 2, we assessed consistency, context, correctness, and detail orientation, as scored by GPT-4V. In Figure 3, we evaluated performance based on Recall, Precision, and F1, which calculate the probability of keyword hits, i.e., the extent to which the generated results cover the keywords.

From Table 2 and Figure 3, we can observe that: 1) Removing the Satellite Expert Database leads to significant performance drops in both tasks. Consistency and context scores fall to 2.78 and 3.00 in Satellite Multi-turn Dialogue, and recall and F1 drop to 0.32 and 0.33, highlighting the importance of expert knowledge. 2) Removing the Retrieval Mechanism results in the lowest scores across most categories. Consistency and correctness for Satellite Multi-turn Dialogue drop to 2.04 and 2.22, and recall and precision for Satellite Visual Question-Answer fall to 0.28 and 0.33, demonstrating the mechanism's crucial role. 3) The variant without the Satellite Rule module exhibits a noticeable decrease in performance,



(a) Sat-RIA on commonsense dialogue. (b) Sat-RIA on visual question-answer.

Figure 4: Result on different LLM models and Sat-RIA on commonsense dialogue and visual question-answer.

particularly in the detail metric for Satellite Visual Question-Answer in Table 2, dropping from 6.61 to 6.16. In Figure 3, the F1 score for Satellite Visual Question-Answer drops from 0.64 to 0.61. This suggests that predefined rules are essential for maintaining the detail and accuracy of generated descriptions. 4) Removing Iterative Alignment lowers performance in both tasks. For Satellite Multi-turn Dialogue, detail drops to 7.65, and F1 drops to 0.50. This demonstrates the importance of iterative alignment in refining content and improving relevance and accuracy. 5) The full model (Ours) consistently achieves the highest scores across all metrics, confirming that combining all modules results in the best performance for both tasks. All these observations demonstrate the effectiveness of each component in our model.

5.7 Analysis on SatDiag

We present examples of results generated by our approach compared to other methods, as shown in Figure 4 (a). Results generated using ChatGPT-4o exhibit hallucinations. Additionally, incorporating our designed commonsense conversation instruction makes the generated results more readable and comprehensible. However, due to the inherent hallucinations in LLMs, the provided reference links either do not exist or do not match the answers. Results generated using the LLaMa-8B model not only exhibit hallucinations but also fail to understand questions in the satellite domain accurately. Incorporating commonsense conversation instruction in this model does not yield satisfactory results either. Our model can generate easily understandable results by integrating satellite knowledge, thereby reducing hallucinations. Moreover, the provided references completely match the answers. By incorporating a KG, the responses are enriched and more comprehensible.

5.8 Analysis on SatVQA

We further present examples of results on Satellite Visual Question-Answer generated by our approach compared to other methods, as shown in Figure 4(b). Results generated using ChatGPT-4 are accurate but often fail to capture the key points, making it difficult for those unfamiliar with satellites to quickly comprehend the information. Conversely, results from the LLava 7B model are not only inaccurate but also challenging for laypersons to understand. Our model, on the other hand, generates results that include object detection and recognition within the images, along with some characteristics of the targets. Additionally, by presenting the results in a simple, easy-to-understand manner, including the use of emojis, comprehension is significantly improved. We also provide relevant satellite links, enhancing the explanatory power and reliability of the generated results.

6 Conclusion

This paper introduces a multimodal, retrieval-augmented framework for satellite commonsense conversation. Our framework integrates a satellite knowledge graph, expert database, and image database to reduce hallucinations and enhance response interpretability. It also features a commonsense conversation instruction to improve answer clarity. The retrieval-augmented iterative alignment module refines responses through multi-stage evaluations aligned with task-specific standards. To evaluate our model, we construct satellite multi-turn dialogue and visual question-answer datasets. Experiments on satellite dialogue and VQA datasets show superior performance over existing LLMs, with fewer hallucinations and more comprehensible answers.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. The corresponding author is Qian Li. The authors of this paper were supported by the NSFC through grant No.62402054, No.62425203 and No.62032003, and the 76th batch of general grants from China Postdoctoral Science Foundation through grant 2024M760279, and supported by the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant Number.

Contribution Statement

Qian Li, Xuchen Li, and Zongyu Chang have made equal contributions and are the co-first authors of this paper.

References

[Bondi *et al.*, 2022] Elizabeth Bondi, Haipeng Chen, Christopher D. Golden, Nikhil Behari, and Milind Tambe. Micronutrient deficiency prediction via publicly available satellite data. In *AAAI*, pages 12454–12460, 2022.

[Carolan *et al.*, 2024] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A review of multi-modal large language and vision models. *CoRR*, 2024.

[Chang *et al.*, 2024] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *TIST*, 15(3):1–45, 2024.

[Chen *et al.*, 2021] Boyo Chen, Buo-Fu Chen, and Yun-Nung Chen. Real-time tropical cyclone intensity estimation by handling temporally heterogeneous satellite data. In *AAAI*, pages 14721–14728, 2021.

[Chen *et al.*, 2023a] Xiao Chen, Ruidan Luo, Ting Liu, Hong Yuan, and Haitao Wu. Satellite navigation signal authentication in GNSS: A survey on technology evolution, status, and perspective for BDS. *Remote. Sens.*, 15(5):1462, 2023.

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

[Choi, 2024] Chang-Sik Choi. Analysis of a delay-tolerant data harvest architecture leveraging low earth orbit satellite networks. *IEEE J. Sel. Areas Commun.*, 42(5):1329–1343, 2024.

[de Witt *et al.*, 2021] Christian Schröder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Alfredo Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. Rainbench: Towards data-driven global precipitation forecasting from satellite imagery. In *AAAI*, pages 14902–14910, 2021.

[Gadiraju *et al.*, 2020] Krishna Karthik Gadiraju, Bharathkumar Ramachandra, Zexi Chen, and Ranga Raju Vatsavai. Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery. In *KDD*, pages 3234–3242, 2020.

[Gallardo, 2024] Isaiah Gallardo. Using reinforcement learning to iteratively construct road networks from satellite images and GPS data. In *AAAI*, pages 23740–23741, 2024.

[Han *et al.*, 2020] Sungwon Han, Donghyun Ahn, Sungwon Park, Jeasurk Yang, Susang Lee, Jihee Kim, Hyunjoo Yang, Sangyoon Park, and Meeyoung Cha. Learning to score economic development from satellite imagery. In *KDD*, pages 2970–2979, 2020.

[Harvel *et al.*, 2024] Nicholas Harvel, Felipe Bivort Haiek, Anupriya Ankolekar, and David Brunner. Can llms answer investment banking questions? using domain-tuned functions to improve LLM performance on knowledge-intensive analytical tasks. In *AAAI*, pages 125–133, 2024.

[Hassanin *et al.*, 2024] Mohammed Hassanin, Marwa Keshk, Sara Salim, Majid Alsubaie, and Dharmendra Sharma. PLLM-CS: pre-trained large language model (LLM) for cyber threat detection in satellite networks. *CoRR*, 2024.

[Huang and Chang, 2023] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *ACL*, pages 1049–1065, 2023.

[Huang *et al.*, 2023] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. In *NeurIPS*, 2023.

[Inkollu and Sastry, 2024] Uma Maheswara Rao Inkollu and J. K. R. Sastry. Ai-driven reinforced optimal cloud resource allocation (ROCRA) for high-speed satellite imagery data processing. *Earth Sci. Informatics*, 17(2):1609–1624, 2024.

[Jablonka *et al.*, 2023] Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, L. Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A. de Jong, Matthew L. Evans, Nicolas Gastellu, Jerome Genzling, María Victoria Gil, Ankur K. Gupta, Zhi Hong, Alishba Imran, Sabine Kruschwitz, Anne Labarre, Jakub Lála, Tao Liu, Steven Ma, Sauradeep Majumdar, Garrett W. Merz, Nicolas Moitessier, Elias Moubarak, Beatriz Mouriño, Brenden Pelkie, Michael Pieler, Mayk Caldas Ramos, Bojana Rankovic, Samuel G. Rodriques, Jacob N. Sanders, Philippe Schwaller, Marcus Schwarting, Jiale Shi, Berend Smit, Ben E. Smith, Joren Van Heck, Christoph Völker, Logan T. Ward, Sean Warren, Benjamin Weiser, Sylvester Zhang, Xiaoqi Zhang, Ghezal Ahmad Zia, Aristana Scourtas, K. J. Schmidt, Ian T. Foster, Andrew D. White, and Ben Blaiszik. 14 examples of how llms can transform materials science and chemistry: A reflection on a large language model hackathon. *CoRR*, 2023.

[Kalyan, 2024] Katikapalli Subramanyam Kalyan. A survey of GPT-3 family large language models including chatgpt and GPT-4. *Nat. Lang. Process. J.*, 6:100048, 2024.

[Kasneci *et al.*, 2023] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[Latif *et al.*, 2024] Ehsan Latif, Ramviyas Parasuraman, and Xiaoming Zhai. Physicsassistant: An llm-powered interactive learning robot for physics lab investigations. *CoRR*, 2024.

[Lebedev *et al.*, 2019] Vadim Lebedev, Vladimir Ivashkin, Irina Rudenko, Alexander Ganshin, Alexander Molchanov, Sergey Ovcharenko, Ruslan Grokhovetskiy, Ivan Bushmarinov, and Dmitry Solomentsev. Precipitation nowcasting with satellite imagery. In *KDD*, pages 2680–2688, 2019.

[Li *et al.*, 2023] Xuechun Li, Paula M. Bürgi, Wei Ma, Hae Young Noh, David Jay Wald, and Susu Xu. Disasternet: Causal bayesian networks with normalizing flows for cascading hazards estimation from satellite imagery. In *KDD*, pages 4391–4403, 2023.

[Li *et al.*, 2024a] Yansheng Li, Bo Dang, Wanchun Li, and Yongjun Zhang. Glh-water: A large-scale dataset for global surface water detection in large-size very-high-resolution satellite imagery. In *AAAI*, pages 22213–22221, 2024.

[Li *et al.*, 2024b] Zhuang Li, Levon Haroutunian, Raj Tumuluri, Phil Cohen, and Gholamreza Haffari. Improving cross-domain low-resource text generation through LLM post-editing: A programmer-interpreter approach. In *EACL*, pages 347–354, 2024.

[Liu *et al.*, 2023a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

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

[Long *et al.*, 2024] Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. Generative multi-modal knowledge retrieval with large language models. In *AAAI*, pages 18733–18741, 2024.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[Lu *et al.*, 2021] Hui Lu, Qi Liu, Xiaodong Liu, and Yonghong Zhang. A survey of semantic construction and application of satellite remote sensing images and data. *J. Organ. End User Comput.*, 33(6):1–20, 2021.

[Lu *et al.*, 2024] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, 2024.

[Segovia-Dominguez *et al.*, 2021] Ignacio Segovia-Dominguez, Huikyo Lee, Yuzhou Chen, Michael J. Garay, Krzysztof M. Gorski, and Yulia R. Gel. Does air quality really impact COVID-19 clinical severity: Coupling NASA satellite datasets with geometric deep learning. In *KDD*, pages 3540–3548, 2021.

[Thirunavukarasu *et al.*, 2023] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[Wang *et al.*, 2003] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[Young *et al.*, 2024] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. *CoRR*, 2024.

[Zhang *et al.*, 2024] Zhengde Zhang, Yiyu Zhang, Haodong Yao, Jianwen Luo, Rui Zhao, Bo Huang, Jiameng Zhao, Yipu Liao, Ke Li, Lina Zhao, Jun Cao, Fazhi Qi, and Changzheng Yuan. Xiwu: A basis flexible and learnable LLM for high energy physics. *CoRR*, 2024.

[Zhao *et al.*, 2024] Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. Alignngpt: Multi-modal large language models with adaptive alignment capability. *CoRR*, 2024.

[Zhou *et al.*, 2023] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. A comprehensive survey on pretrained foundation models: A history from BERT to chatgpt. *CoRR*, 2023.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, 2023.