Advancement in Graph Understanding: A Multimodal Benchmark and Fine-Tuning of Vision-Language Models

Qihang Ai[1*], **Jiafan Li**[2,3*], **Jincheng Dai**[4], **Jianwu Zhou**[1], **Lemao Liu**[6], **Haiyun Jiang**[5†], **Shuming Shi**[7]
[1]Beijing Institute of Technology
[2]Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Zhejiang University [5]Sun Yat-sen University [6]Wechat AI, Tencent, China [7]Tencent AI Lab, China
qihang-ai@bit.edu.cn, lijiafan23@mails.ucas.ac.cn, jincheng_dai@zju.edu.cn,
1120211477@bit.edu.cn, redmondliu@tencent.com, jianghy66@mail.sysu.edu.cn,
shumingshi@tencent.com

Abstract

Graph data organizes complex relationships and interactions between objects, facilitating advanced analysis and decision-making across different fields. In this paper, we propose a new paradigm for interactive and instructional graph data understanding and reasoning. Instead of adopting complex graph neural models or heuristic graph-to-text instruction design, we leverage Vision-Language Models (VLMs) to encode the graph images with varying structures across different domains. This paper first evaluates the capabilities of public VLMs in graph learning from multiple aspects. Then it introduces a novel instruction-following dataset for multimodal graph understanding and reasoning in English and Chinese. Besides, by fine-tuning MiniGPT-4 and LLaVA on our dataset, we achieved an accuracy increase of 5%-15% compared to baseline models, with the best-performing model attaining scores comparable to Gemini in GPT-asissted Evaluation. This research not only showcases the potential of integrating VLMs with graph data but also opens new avenues for advancement in graph data understanding.

1 Introduction

Graph is an important form of structured data, which is capable of storing and representing the complex relationships between objects. Downstream tasks performed on graph data include node classification (Xiao et al., 2022), link prediction (Zhang and Chen, 2018), graph reasoning (Chen et al., 2020) etc. In early days, graph learning is commonly modelled using graph networks, with a wealth of seminal literature supporting this approach. In recent years, graph neural networks(Li et al., 2015; Dai et al., 2018; Battaglia et al., 2018; Fan et al., 2019; Zhang and Chen, 2018) provide a more flexible and effective means of dealing with the diversity and complexity in graph structures. By propagating and aggregating node information within the graph structure, GNNs capture complex graph features.

With the rise of Large Language Models (LLMs), there has been extensive research in the field of *natural language-guided interactive graph data understanding*, showcasing two main strategies. The first approach constructs prompts by explicitly representing graph structures in a sequential format. Constructing prompts for LLMs involves innovative techniques such as self-prompting (Guo et al., 2023), graph-syntax trees (Zhao et al., 2023b), natural language descriptions of graphs' structures and features (Ye et al., 2023b), and graph-structure prompting in various modalities (Das et al., 2023). The second strategy transforms GNN-learned features into LLM-comprehensible tokens. Models like GIT-Former (Liu et al., 2024a) GIMLET (Zhao et al., 2023a) and MolCA (Liu et al., 2024b) consider graph as a mode and integrate all modality data into a unified latent space.

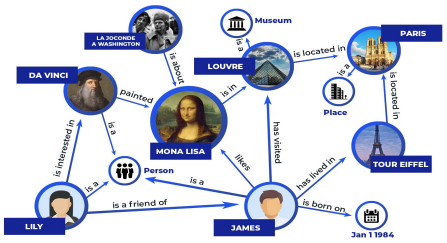
However, due to the inherent differences between graph structures and language models, current integration methods face significant limitations. First, sequence-based prompt representations inevitably lose crucial graph structural information, which is vital for complex graph understanding and reasoning(Ge et al., 2023; Li et al., 2024). Second, integrating GNN-learned features into LLMs presents alignment challenges, as the representation learning space of GNNs cannot be easily mapped onto the token space of LLMs through simple function mappings(Xue et al., 2023).

Considering these limitations, this work introduces a new paradigm for graph data understanding based on VLMs. Our basic idea is to transform graph data into images, either through their natural representations like maps or visualization

[*]Equal contribution
[†]Corresponding author(jianghy66@mail.sysu.edu.cn)

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7485–7501
August 11-16, 2024

methods[†]. This is followed by utilizing an image encoder to interpret the visual information, thereby understanding the semantic information of nodes and relationships within the graph data. The final step of this process involves the fusion of image and text encoders, which is inherently supported in specific VLMs (Kim et al., 2021; Li et al., 2022a; Liu et al., 2023), offering a novel approach to graph data comprehension. Figure 1 shows an example of graph data access based on visual language models.



Instruction:
What is the relationship between Da Vinci and the Mona Lisa?
Response:
Da Vinci painted the Mona Lisa.

Figure 1: Graph understanding with the paradigm of instruction following by vision-language models.

Utilizing visual language models for graph data comprehension offers friendly and natural interaction, reducing manual labor costs, and unifies diverse graph structures globally, handling various graph-related tasks through natural language instructions. Leveraging state-of-the-art visual language models trained on extensive datasets facilitates knowledge transfer and promises high performance potential in this paradigm of graph understanding and reasoning.

This work introduces a novel dataset on various graph formats for multimodal graph understanding and reasoning in English and Chinese. Instructions involve simple queries and complex multi-hop reasoning on the graphs, with difficulty levels varying from simple to medium and difficult. Additionally, based on this dataset, we conducted a comprehensive evaluation of current open models from various perspectives. In conclusion, the current models exhibit significantly better performance in English compared to Chinese. Even the most powerful models like Genimi or GPT-4V achieve accuracy rates below 30%. Finally, in fine-tuning LLaVA and MiniGPT-4 on our dataset, experimental results revealed that the fine-tuned LLaVA model on English&Chinese datasets outperformed baseline LLaVA on both English and Chinese dataset and achieved comparable performance with Genimi in Chinese. To enhance the open-source models' ability in recognizing Chinese characters, we also proposed incorporating OCR data transformed into instruction data into our Chinese datasets. The model refined through this approach exhibited significant improvements on the Chinese dataset after fine-tuning.

Our main contributions are as follows:

- We constructed a dataset for multimodal graph understanding and reasoning, providing instructions for graph-related questions and corresponding ground truth answers.
- We evaluated the capabilities of public VLMs in graph learning, analyzing their performance and limitations across various graph tasks.
- We fine-tuned the open-source models LLaVA and MiniGPT-4 using the constructed dataset, resulting in significant improvements on both Chinese and English datasets.

2 Related Work

2.1 Integrate LLMs with Graph Learning

With the emergence of LLMs, the applications in graph-related tasks have surpassed traditional GNN-based methods. Currently, there are three main ways of integrating large language models with graph data.

LLMs-as-Enhancers. This research line aims to enrich the node attribute and relation representations using the capabilities of LLMs. For example, TAPE (He et al., 2023) leverages the knowledge of large language models to generate high-quality node features, thereby enhancing the quality of initial node embeddings in GNNs. Knowledge-Enhanced Augmentation (KEA) (Chen et al., 2024) enriches text attributes by providing additional information, stimulating LLMs to generate lists of knowledge entities and their descriptions and encoding them through fine-tuned PLMs and deep sentence embedding models.

LLMs-as-Predictors. This method represents graph node attributes and structures in the form of prompts, using LLMs to directly generate predictive patterns. For instance, InstructGLM (Ye et al., 2023b)replaces GNN's predictors with LLMs.

[†]https://www.ownthink.com/

	# Train	# Test	# English	# Chinese	Overall
Knowledge Graph	1644	413	986	1071	2057
Route Map	1616	406	1071	951	2022
Flowchart	636	160	714	82	796
Mind Map	950	238	594	594	1188
Gantt Chart	475	138	564	49	613
Overall	5321	1355	3929	2747	6676

Table 1: An overview of our multimodal instruction-following benchmark on graph data. "#" means the number of instruction-response pairs. dataset-stat.

GPT4Graph (Guo et al., 2023) adopts graph description language of prompt engineering, improving collaborative working methods in various situations. GraphGPT (Tang et al., 2023) aligns LLMs with graph structural knowledge through a graph-guided tuning paradigm.

LLM and Graph Collaboration. Aligning the embedding spaces of graph models and LLMs achieves the integration of graph and text modalities. Text2Mol (Edwards et al., 2021) proposes a cross-modal attention mechanism using transformer decoder for early fusion of graph and text embeddings. "Think on Graph" (Sun et al., 2024) presents the "LLM $\otimes$ KG" paradigm, a new approach integrating LLMs and knowledge graphs (KG). It treats large language models as agents for interactively exploring entities and relations in KGs. RLMRec (Ren et al., 2023) suggests aligning the semantic space of LLMs with the representation space of collaborative relational signals in recommendation systems through contrastive modeling.

2.2 Multimodal Large Language Models

Recently, with the rapid advancement of large language models and their demonstrated powerful interactive capabilities, a new paradigm has been proposed for the vision-language tasks. Based on an encoder-decoder framework and utilizing LLMs as decoders (Liu et al., 2023; Li et al., 2023; Wang et al., 2023; Su et al., 2023), Multimodal Large Language Models (MLLMs) exhibit significant multimodal capabilities across various benchmarks. This approach leverages cross-modal transfer, enabling the sharing of knowledge between language and multimodal domains (Zhu et al., 2023).

Visual instruction tuning (Liu et al., 2023) is employed to develop a MLLM that is adept at general-purpose visual and language understanding. LLaVA (Liu et al., 2023) extends the self-instruction (Wang et al., 2022) approach to the multimodal field by translating images into texts with captions and bounding boxes. Besides LLaVA, many other powerful MLLMs have also emerged built upon LLMs (Yin et al., 2023), including the open-source ones, e.g., MiniGPT-4 (Zhu et al., 2023), mPLUG-Owl2 (Ye et al., 2023a), Multimodal-GPT (Gong et al., 2023), and the commercial models, e.g., GPT-4V (OpenAI, 2023) and Gemini (Team et al., 2023).

3 Dataset

We constructed two multimodal instruction-following datasets on graphs in English and Chinese respectively. Each dataset includes five types of graphs: knowledge graph, route map, mind map, flowchart, and Gantt chart. The images were crawled from search engines and filtered for relevance and accuracy by humans. The construction of multimodal instructions involves two steps. The first step prompts GPT-4V to generate candidate instructions and responses for each image. The second step involves human evaluation and annotation to ensure the validation of instructions and accuracy of responses. For images containing mixed language content, such as English text in Chinese images or vice versa, we have already discarded these during the manual filtering process to ensure language consistency in our dataset.

3.1 Data Annotation

After manually filtering out unclear and irrelevant images which don't belong to any graph type[†], We obtained a total of 2,807 images. Among these, 1,493 are Chinese images, including 517 knowledge graphs, 335 mind maps, 584 route maps, 49 flowcharts, and 8 Gantt charts. The 1,314 English

[†]Images containing mixed language content, such as English in Chinese images or vice versa, were discarded.

Valid rate	GPT-4V	Gemini	LLaVA	MiniGPT-4
Chinese	30.6	**36.0**	20.0	0
English	**97.2**	88.9	88.5	73.0

Table 2: The VLMs' ability to generate instructions in response to Chinese and English images. The numbers in the table represent the valid rate for Chinese images or English images in percentage form. The best results are identified with **bold**. ins-generation.

images are comprised of 251 flowcharts, 208 Gantt charts, 257 knowledge graphs, 216 mind maps, and 382 route maps. The proportion of Chinese to English annotated images is approximately 1.14:1.

Given a valid image, we first annotated it by GPT-4V under our elaborated prompts as shown in Figure 14, during which Chinese images were annotated using prompts in Chinese and English images were annotated with English prompts. For each image, under our prompt, GPT-4V usually posed three questions: simple, medium, and complex, as shown in Figure 19 and Figure 20 in Appendix A.4.

Following this automatic annotation process, we conducted an evaluation of the relevance of the instructions generated by GPT-4V in relation to the images. This critical analysis aimed to ascertain the extent to which GPT-4V's output aligned with the images. In our in-depth examination, the instructions generated by GPT-4V might be invalid due to either of the following two aspect:

- Instructions are completely unrelated to the image's content.
- Instructions are related to the image's theme but involving nodes or edges not present in the image.

For each image with several instruction-response pairs, we then implemented a manual fine-annotation strategy. Valid instructions were retained while invalid ones were discarded. Subsequently, we checked the correctness of responses corresponding to the valid instructions and corrected any incorrect responses. In cases where all instructions for an image were deemed invalid, we manually supplemented two instruction-response pairs: one simple instruction and one complex instruction. This approach ensures a high standard of instruction-response pairs in our benchmarks.

3.2 Data Statistics and Analysis

Table 1 presents the statistical information of our multimodal instruction-following benchmark dataset. All dataset construction processes undergo rigorous manual evaluation to ensure high quality. To demonstrate this, we asked an unseen annotator to randomly select a sample of 50 Chinese and English samples (i.e., image, instruction, response), respectively. Each image was subjected to a rigorous quality assessment within the selected samples. Our analysis revealed that out of the 100 samples, an impressive 97 were classified as qualified, highlighting the dataset's high standard of clarity and uniqueness.

3.3 Evaluation Protocols

To facilitate the evaluation of model performance using our graph vision-language dataset, we provided two evaluation protocols. The first is a manual assessment of model accuracy, where human evaluators were asked to determine whether the model response satisfies the instruction conditioned on an image (Zhu et al., 2023). The second evaluation protocol we used is a GPT-assisted visual instruction assessment. Building on a previous work (Liu et al., 2023) that employed GPT-4 (text-only) to score answers by comparing the output of GPT-4 with that of another model, we have adopted a similar approach for GPT-4V evaluation and designed an English prompt shown in Figure 16 and a Chinese prompt in Figure 17 in Appendix A.3.

4 Evaluation on Public Vision-Language Models

4.1 What Abilities Do We Focus On?

The initial phase of our study involved a manual evaluation of GPT-4V (OpenAI, 2023), Gemini (Team et al., 2023), LLaVA (Liu et al., 2023), and MiniGPT-4 (Zhu et al., 2023)'s capability to generate and follow instructions in response to images.

Instruction Generation Ability. Given an image of graph, we first evaluate the capability of VLMs on generating graph-related instructions. We determined the validity of each instruction as described in Section 4.3.

Model	Simple	Medium	Complex	Multi-hop	Information noise	Dense	Sparse
Results of response accuracy for Chinese images.							
MiniGPT-4	8	0	0	0	0	0	4
LLaVA	32	12	0	0	0	16	20
Gemini	**60**	24	12	4	4	20	36
GPT-4V	56	**40**	**28**	**20**	**8**	**48**	**52**
Results of response accuracy for English images.							
MiniGPT-4	36	24	32	20	28	8	48
LLaVA	52	20	40	52	0	24	52
Gemini	**80**	64	68	**76**	28	24	52
GPT-4V	76	**80**	**72**	64	**56**	**64**	**88**

Table 3: Results of response accuracy under different types of instructions. Dense and sparse represent the information density of images. Simple, medium and complex represent the difficulty level of instructions. The numbers in the table represent the accuracy rate in percentage form. The best results are identified with **bold**. instruction-eval.

Various Instruction Following Abilities. Given a valid instruction of an image, we also evaluated the correctness of the response generated by existing VLMs. We assessed various ability dimensions, including simple and intermediate-level instruction following ability, complex instruction following ability, multi-hop reasoning ability, robustness to noise, performance across different information densities, and an ablation study on different types of graphs. These dimensions were explored by providing corresponding instructions (see Section 4.4, 4.5, 4.6, 4.7, 4.8, 4.9).

4.2 Evaluation Settings

Instruction Validity. We sampled a set of 50 images for the experiment, encompassing five types of graphs with a balanced distribution between English and Chinese. The prompts we used are shown in Figure 14 in Appendix A.3. As mentioned in Section 3.1, there may exist two types of invalid instructions. The distribution of valid rate for different models is detailed in Table 2.

Response to Various Instructions. For each ability dimension, we randomly sampled 25 instructions for evaluation from our dataset. To enhance the quality of responses, we utilized pre-established prompts as detailed in Figure 15 in Appendix A.3.

The evaluation metric focused on the accuracy of the VLMs' responses to the instructions. This was determined by human evaluators. To provide a comprehensive understanding, we separately evaluated the performance of VLMs on images with content in Chinese and English [†]. The assessment result is presented in Table 3.

4.3 Instruction Generation Ability

During the Chinese image instruction generation, we observed low performance for all the models, with less than 40% valid rate. The main reason for the low performance was the misrecognition of node names. Specifically they struggled to correctly identify the Chinese characters within them. As a result, these models tend to generate erroneous node and relation names based on their internal knowledge. We illustrate this hallucination by an example in Figure 21 in Appendix A.4.

4.4 Simple and Medium Instruction Following Ability

As mentioned in Section 3.1, the instructions of our dataset were initially generated by GPT-4V, where each sample was additionally labeled with *simple, medium*, or *complex*. The accuracy results for simple and medium instructions are shown in Table 3.

MiniGPT-4 lacked basic cognition in discerning graph types and was unable to differentiate between types of graphs. A common challenge faced by all these models, including GPT-4V, Gemini, LLaVA, and MiniGPT-4, was their struggle with identifying the number of nodes in graph-type data.

4.5 Complex Instruction Following Ability

Complex instruction following tasks require understanding intricate commands but may result in simple answers. The corresponding evaluation results are shown in Table 3.

[†]In cases where images contain both Chinese and English content, classification as Chinese or English was determined based on the predominant language used in the nodes and edges of the graph.

Model	Knowledge Graph	Mind Map	Route Map	Flowchart	Gantt Chart	Overall
Results of response accuracy for Chinese images.						
MiniGPT-4	8	0	4	4	0	3.2
LLaVA	20	12	20	20	8	16
Gemini	24	**44**	24	20	**40**	29.6
GPT-4V	**32**	40	**32**	**36**	28	**32**
Results of response accuracy for English images.						
MiniGPT-4	64	16	8	40	24	30.4
LLaVA	56	28	16	44	32	35.2
Gemini	76	72	60	**80**	40	65.6
GPT-4V	**96**	**80**	**76**	72	**42**	**73.2**

Table 4: Results of response accuracy under different types of graphs. Knowledge graph, mind map, route map, flowchant, and Gantt chart represent the type of graph in the image. The numbers in the table represent the accuracy rate in percentage form. The best results are identified with **bold**. graph-eval.

When processing images containing English content, the performance of Gemini, LLaVA, and MiniGPT-4 in complex instructions surpasses their accuracy in medium instructions. This improved performance can be attributed to the nature of complex instructions, which some encompass open-ended questions relying on coarse-grained visual information. Such scenarios play to the strengths of LLMs, leveraging their robust linguistic capabilities to effectively interpret and respond to these complex instructions.

4.6 Multi-hop Reasoning Ability

As mentioned in Section 4.5, complex instructions may result in simple answers. In contrast, multi-hop reasoning instructions involve linking several logical steps, usually leading to more elaborate responses. This posed a significant challenge for VLMs in providing completely accurate responses to multi-hop reasoning tasks. The corresponding evaluation results are shown in Table 3.

4.7 Robustness to Noise

As mentioned in Section 3.1, we have excluded blurry images from our dataset. In this part, we have additionally collected a set of noisy images, comprising 25 instructions each for both Chinese and English content. The corresponding results are shown in Table 3.

MiniGPT-4 occasionally struggled with noisy images, which was evident in instances of language confusion, such as responding in English to prompts and images that are in Chinese, and producing meaningless repetitive answers. GPT-4V tended to offer vague responses or guiding suggestions, stating that due to image quality limitations, it couldn't give valid responses.

4.8 Performance across Different Information Densities

Different images can contain varying content and details, i.e., different information densities. Information density, typically judged by the number of nodes and edges in a graph, ranged from sparse to dense and is assessed manually. This section explores how the VLMs perform with images of varying information densities. The corresponding evaluation results are shown in Table 3 and GPT-4V outperformed other VLMs in processing image information densities in both English and Chinese contexts.

4.9 Ablation Study on Different Types of Graphs

To comprehensively assess how the VLMs perform across various types of graphs, we randomly sampled 25 instructions for each graph type, with the related results being displayed in Table 4.

In the analysis of images containing English content, all four VLMs displayed a relative familiarity with the structure of knowledge graphs, achieving their highest accuracy in this category. In route maps, GPT-4V not only led in accuracy for both Chinese and English content images but also provided more precise and standardized responses, using both absolute directions(north, south, east, west) and relative directions(up, down, left, right). In contrast, Gemini used only absolute directions, while LLaVA and MiniGPT-4 depended entirely on relative directions.

Model	Manual Evaluation			GPT-assisted Evaluation		
	English	Chinese	English&Chinese	English	Chinese	English&Chinese
Gemini	**82**	36	53	6.42	4.12	5.15
GPT-4V	74	**44**	**59**	**8.66**	**6.66**	**7.55**
MiniGPT-4	14	1	5	1.92	1.13	1.38
LLaVA	40	13	25	4.20	2.47	3.59
MiniGPT-4 +English	22(+8)	-	-	2.17(+0.25)	-	-
MiniGPT-4 +Chinese	-	5(+4)	-	-	1.42(+0.29)	-
MiniGPT-4 +English&Chinese	-	-	13(+8)	-	-	2.32(+0.94)
LLaVA +English	55(+15)	-	-	6.03(+1.83)	-	-
LLaVA +Chinese	-	27(+14)	-	-	3.34(+0.87)	-
LLaVA +English&Chinese	-	-	40(+15)	-	-	4.91(+1.32)

Table 5: Fine-tuned model results. The numbers in the table under the "Manual Evaluation" column represent the accuracy rate in percentage form. In the "GPT-assisted Evaluation" column, the numbers represent the scores given to each model by GPT-4V. Each model receives an overall score on a scale of 1 to 10, based on a comparison between the responses generated by the model and the ground truth answers, along with an accompanying explanation. result.

5 Experiment

5.1 Baseline Models and Settings

MiniGPT-4 consists of a vision encoder with a pretrained ViT (Dosovitskiy et al., 2021) and Q-Former (Li et al., 2022b), a single linear projection layer, and an advanced Vicuna (Chiang et al., 2023) large language model. During the training process, only the parameters of the Q-Former and linear projection layers are fine-tuned, while the parameters of the language and visual models are kept unchanged.

LLaVA uses language-only GPT-4 to generate multimodal language-image instruction-following data. This approach allows LLaVA to connect a vision encoder and a language model for general-purpose applications. The fine-tuning stage involved updating both the pre-trained weights of the projection layer and LLM in LLaVA, while keeping the vision encoder fixed.

Both LLaVA and MiniGPT-4 were trained for 10 epochs. Performance was evaluated after all the intervals, and the model demonstrating the best performance was then selected for data generation.

Experiment Details We divided the training set and the test set according to the ratio of 4:1. Specific data is shown in Table 1. We used the English training sets, the Chinese training sets and the combination of Chinese and English training sets for fine tuning, respectively.

5.2 Results

5.2.1 Qualitative Analysis

After fine-tuning, the model demonstrates many advanced abilities compared to the baseline model. Here, we will analyze and explain in detail with specific examples based on the best-performing model, named LLaVA FT. These cases are shown in Appendix A.1.

In Figure 2, LLaVA FT effectively utilizes visual cues to discern pertinent information within images, enabling accurate interpretation of depicted processes. In constrast, the baseline LLaVA model relies solely on contextual understanding, thereby neglecting crucial image details in its responses. The scenario depicted in Figure 3 highlights LLaVA FT's proficiency in accurately arranging tasks based on their sequence in Gantt charts, resulting in precise identifications. Conversely, the baseline LLaVA model struggles to precisely identify the sequence and frequently misinterprets presented tasks. And Figure 4 demonstrates LLaVA FT's proficiency in identifying and categorizing relationships within a knowledge graph, while Figure 5 showcases its enhanced ability to determine the quantity of nodes within a given graph, providing specific and relevant responses.

These cases demonstrate the enhanced capabilities of the fine-tuned model in understanding flowcharts, Gantt charts, knowledge graphs, and mind maps. The fine-tuned model's ability to answer questions based on graphical data in both English and Chinese has been significantly improved.

5.2.2 Quantitative Analysis

We used the evaluation method described in Section 3.3. In Manual Evaluation, we randomly sampled 100 instructions from the test sets. Human evaluators assessed the correctness and reasonableness of the model-generated responses. In GPT-assisted Evaluation, we calculated the average value of GPT-

Model	Manual Evaluation		
	English	Chinese	English&Chinese
LLaVA+Chinese	-	27	-
LLaVA+Chinese w OCR	-	**30**	-
LLaVA+English&Chinese	50	28	40
LLaVA+English&Chinese w OCR	**68**	24	**46**
LLaVA+English&Chinese w Multimodal CoT	58	19	41

Table 6: Ablation results. The numbers in the table represent the accuracy rate in percentage form.ablation.

4V scores for each model to measure the effect of fine-tuning models.The results were compared with the baseline and are presented in Table 5.

After fine-tuning, MiniGPT-4 improved by 8%, while LLaVA improved by 15% over the baseline. Mixed-language fine-tuning in both Chinese and English showed similar effectiveness to single-language fine-tuning. The most effective model, LLaVA+English&Chinese, achieved significant improvements, with 28% accuracy on the Chinese test set, 54% on the English test set, and 39.4% overall. Particularly noteworthy is its performance on the Chinese test set, matching that of Gemini.

5.3 Ablation Study

5.3.1 OCR Instructions

In our evaluation in Section 4.4, both MiniGPT-4 and LLaVA exhibited a significant decrease in accuracy when providing responses to Chinese images. To enhance their ability to recognize Chinese characters in image datasets, we have devised targeted OCR instructions. We utilized the training set provided in the Chinese Scene Text Recognition Technology Innovation Competition.[†] The training dataset consists of 212,023 images containing textual information. We tokenized OCR text from images and selected 2,224 closely aligned images for fine-tuning. We created the instruction set by using the prompt "Please recognize the text in the image" and pairing it with the text recognition results from the training dataset. The prompts we used are shown in Figure 18 in Appendix A.3. Subsequently, we combined the OCR instructions with Chinese fine-tuning instructions, adhering to a data augmentation model, to enhance the OCR capabilities. We separately added the OCR instruction to the Chinese fine-tuning dataset and the English&Chinese fine-tuning dataset.

[†]https://aistudio.baidu.com/datasetdetail/8429

5.3.2 CoT Augumented

In our model, we generated a series of intermediate reasoning steps, referred to as Chain-of-Thought(CoT), to enhance the capability for complex reasoning problems. For complex questions, it is necessary to integrate information from both text and images for reasoning, so we used Multimodal CoT (Zhang et al., 2024), which consists of two stages: Rationale Generation and Answer Inference. Initially, caption text and visual features are utilized to obtain a more accurate Rationale (R). Subsequently, the final Answer is derived from R, along with text and visual features.

Our Ablation results are presented in Table 6. There was a noticeable improvement in both the Chinese test set and the Chinese-English mixed test set after incorporating OCR fine-tuning. Particularly, on the mixed-language test set, the performance increased to 46%, marking a substantial improvement of 21% compared to the baseline LLaVA model. The multimodal CoT approach resuls in some improvement compared to the original model. However, there is still a need to design a more suitable multimodal fusion method tailored for graph-related problems, involving the identification of nodes and relationships within the graph.

6 Conclusion and Future Work

This study focuses on natural language-guided interactive graph data understanding, distinguishing itself from traditional graph network modeling by utilizing VLMs to enable flexible interactions with graph data. We constructed a benchmark dataset to address the lack of available data in this domain, further supporting the development and evaluation of models in this field. By fine-tuning LLaVA and MiniGPT-4 with our dataset, we observed notable improvements in performance on both Chinese and English datasets, marking a significant advancement in the field of graph data understanding. In the future, we will advance this work from the following two directions. (1) Designing dedicated im-

age encoders to capture the pixels related to nodes and edges more sensitively. (2) Supporting large graph (with thousands of nodes or edges) understanding by splitting a big image into a sequence of sub-images. It requires that the VLMs have the ability to understand the internal content in a single image as well as the contents across different sub-images, for a more accurate understanding of the original large graph content.

7 Limitations

7.1 Imbalanced Dataset

While there is an abundance of available graphic data resources, obtaining them in the form of images can be particularly challenging. In our constructed dataset, the samples for Gantt charts and flowcharts are notably scarce. This scarcity is primarily due to the limited availability of these specific types of images on the internet, resulting in fewer instructions in our dataset compared to the other three graph types. Despite our efforts to manually filter out noisy images, as highlighted in Section 3.2, where we found that out of 100 randomly sampled images, an impressive 97 were deemed of high quality, our dataset may still contain images of duplication and blurriness. Hence, there's a necessity to explore more reliable automated data cleaning pipelines.

7.2 Hallucination

Our model is built upon LLMs and inherits its limitations. It may suffer from illusions when faced with non-existent knowledge. When querying nodes and relationships in images, the model may respond with nodes not present in the image, relying on its own common sense for answers. As shown in Figure 21 in Appendix A.4, this example demonstrates the model's hallucination in interpreting Chinese instructions, where it incorrectly recognizes "Apply for a refund" as "Select a product" .

References

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org.

Hanjun Dai, Zornitsa Kozareva, Bo Dai, Alex Smola, and Le Song. 2018. Learning steady-states of iterative algorithms over graphs. In *International conference on machine learning*, pages 1106–1114. PMLR.

Debarati Das, Ishaan Gupta, Jaideep Srivastava, and Dongyeop Kang. 2023. Which modality should i use – text, motif, or image? : Understanding graphs with large language models.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The world wide web conference*, pages 417–426.

Qingqing Ge, Zeyuan Zhao, Yiding Liu, Anfeng Cheng, Xiang Li, Shuaiqiang Wang, and Dawei Yin. 2023. Enhancing graph neural networks with structure-based prompt.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. 2023. Gpt4graph: Can large language models understand graph structured data ? an empirical evaluation and benchmarking.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2023. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2024. A survey of graph meets large language model: Progress and future directions.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*. Oral presentation at NeurIPS'23.

Pengfei Liu, Yiming Ren, and Zhixiang Ren. 2024a. Git-mol: A multi-modal large language model for molecular science with graph, image, and text.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2024b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter.

OpenAI. 2023. Gpt-4v(ision) system card. Accessed: 2024-02-01.

Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2023. Representation learning with large language models for recommendation.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2023. Graphgpt: Graph instruction tuning for large language models.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Shunxin Xiao, Shiping Wang, Yuanfei Dai, and Wenzhong Guo. 2022. Graph neural networks in node classification: survey and evaluation. *Machine Vision and Applications*, 33:1–19.

Rui Xue, Xipeng Shen, Ruozhou Yu, and Xiaorui Liu. 2023. Efficient large language models fine-tuning on graphs.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023a. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. 2023b. Natural language is all a graph needs.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023a. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning.

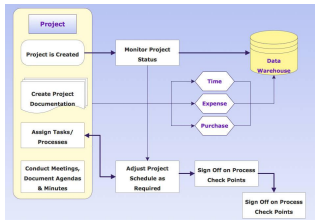
Jianan Zhao, Le Zhuo, Yikang Shen, Meng Qu, Kai Liu, Michael Bronstein, Zhaocheng Zhu, and Jian Tang. 2023b. Graphtext: Graph reasoning in text space.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 More Qualitative Results

Here we list some qualitative results. Yellow highlights correct but irrelevant information. Green highlights the correct response and Red highlights the wrong response.



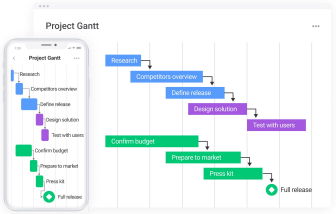
Input Image

Instruction: After creating project documentation, what is the next step in the flowchart?

LLaVA: In the flowchart, the next step after creating project documentation is to create a project schedule. This involves setting up a timeline for the project, including milestones, deadlines, and dependencies between tasks. The project schedule helps ensure that all tasks are completed on time and that the project stays on track.

LLaVA FT: The next step is to "Assign Tasks/Processes".

Figure 2: Flowchart in English



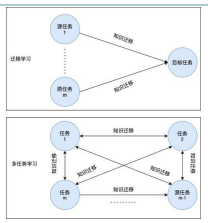
Input Image

Instruction: What is the first task listed in the Gantt chart?

LLaVA: The first task listed in the Gantt chart is "Design Solution".

LLaVA FT: The first task listed is "Research".

Figure 3: Gantt Chart in English



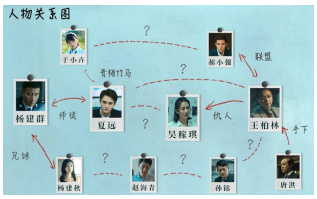
Input Image

Instruction: 图中“源节点1”通过什么关系连接到“目标节点”？
Translation: What is the relationship between "source node 1" and "target node" in the diagram?

LLaVA: “源节点1”通过“节点2”连接到“目标节点”。这个图是一个网络图，其中每个节点代表一个实体或事件，而连接线表示它们之间的关系。
Translation: The "source node 1" is connected to the "target node" through "node 2". The graph is a network diagram, where each node represents an entity or event, and the connecting lines represent the relationships between them.

LLaVA FT: “源节点1”通过“知识迁移”关系连接到“目标节点”。
Translation: The "source node 1" is connected to the "target node" through a "knowledge transfer" relationship.

Figure 4: Knowledge Graph in Chinese



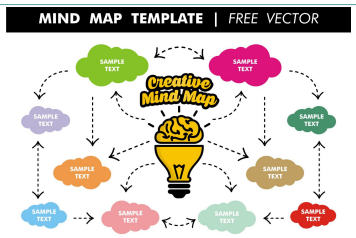
Input Image

Instruction: 图中有几个人物被连接起来了？
Translation: How many characters are connected in the picture?

LLaVA: 在图中，有几个人物被连接起来了。这些人物包括一个男性和一个女性，他们被连接在一起，并且还有其他人物被连接在他们周围。
Translation: In the picture, there are several figures that are connected. These figures include a male and a female who are connected together and there are other figures connected around them.

LLaVA FT: 图中有十个人物被连接起来了。
Translation: There are ten figures in the picture that are connected.

Figure 5: Mind Map in Chinese



Input Image

Instruction: What is the central idea represented in this mind map?

LLaVA: The central idea represented in this mind map is creativity. The image features a yellow light bulb, which symbolizes creativity, at the center of the map, surrounded by various text boxes and arrows.

LLaVA FT: The central idea represented in this mind map is "Creative Mind Map".

Figure 6: Mind Map in English

A.2 More Quantitative Results

A.2.1 Manual Evaluation

Here we list some quantitative results. Yellow highlights correct but irrelevant information. Green highlights the correct response and Red highlights the wrong response.

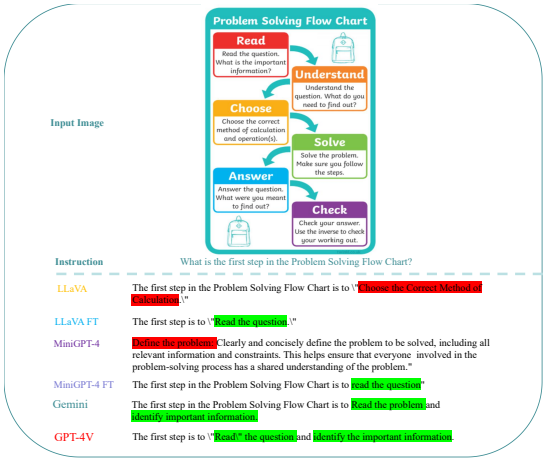


Figure 7: Fine-tuned model results on English&Chinese dataset

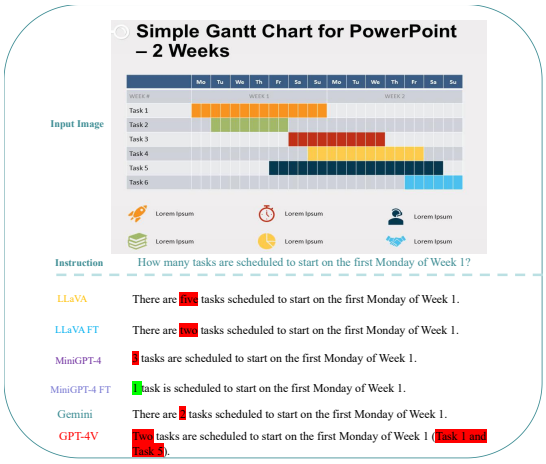


Figure 8: Fine-tuned model results on English dataset

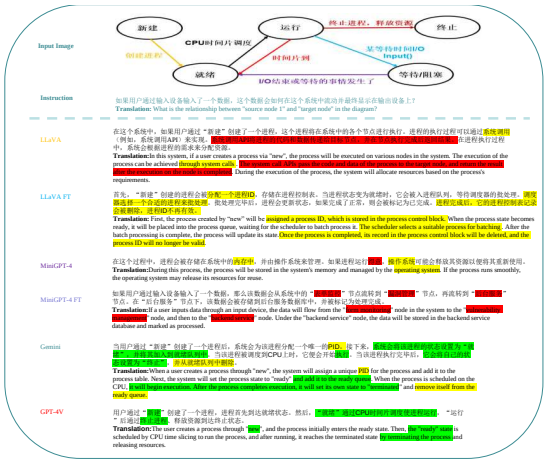


Figure 9: Fine-tuned model results on Chinese dataset

A.2.2 GPT-assisted Evaluation

Here, we list some GPT-assisted score results of the VLMs, including the best-performing fine-tuned models.

Figure 10: Average Scores

Figure 11: Median Scores

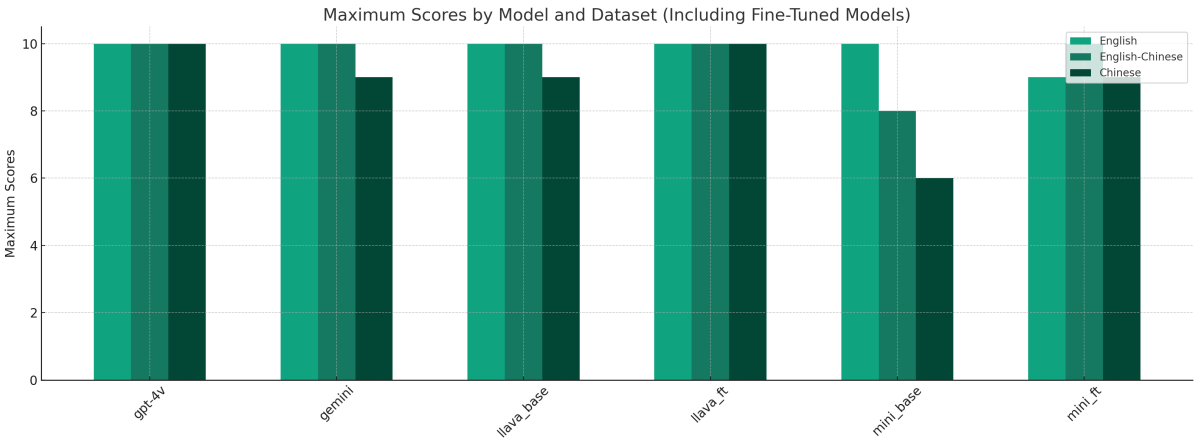


Figure 12: Maximum Scores

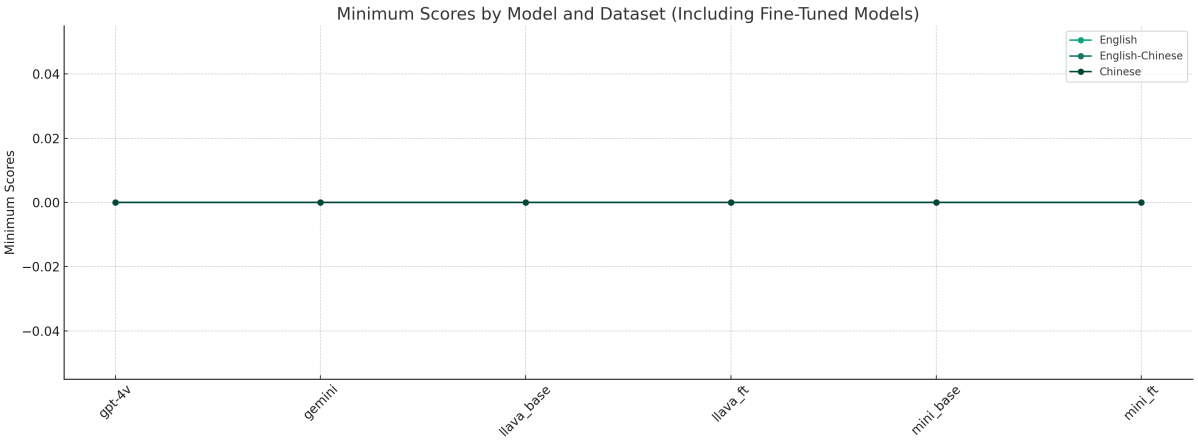


Figure 13: Minimum Scores

A.3 Prompts

The prompt used to generate image-based conversation from GPT-4V is shown in Figure 14. Under this prompt, GPT-4V usually posed three questions for each image: *simple, medium, or complex*.

Prompt Template

English Prompt

You are an expert in the field of **knowledge graph**. You must ensure that the image is correctly identified, and that the questions raised and answers are correct. Please ask and answer some questions based on this image. The questions should preferably be based on specific nodes/edges of the image and should range from simple to complex. The answer should be given in the format of Simple: Question: Answer: Medium:... Complex:...

Chinese Prompt

你是**知识图谱**方面的专家，你一定要保证识别图片正确，提出且回答问题正确。请根据这张图提出且回答一些问题，最好基于该图的具体节点/边来提问，并且提出的问题从简单到复杂，回答按照`Simple: Question: Answer: Medium:... Complex:...`的格式给出。

Figure 14: Example prompt for guiding GPT-4V to create image-based instruction-following data, featuring three levels of complexity: simple, medium, and complex. The **bond** indicates the type of graph, which can be replaced by a flowchart, mind map, Gantt chart, or route map. Yellow highlights the expert identification. Green highlights ensuring correctness and Red highlights the response form.

The prompt used for prompting VLMs to answes the instructions based on an given image is showed in Figure 15.

Prompt For Responding to Images

English Prompt

You are an expert in the field of **knowledge graph**. You must ensure that the image is correctly identified, and that the answers are correct. Please answer some questions based on this image.

Chinese Prompt

你是**知识图谱**方面的专家，你一定要保证识别图片正确，且回答问题正确。请根据这张图回答一些问题。

Figure 15: English and Chinese prompts for guiding VLMs to give responses to instructions based on a given image. The **bond** indicates the type of graph, which can be replaced by a flowchart, mind map, Gantt chart, or route map. Yellow highlights the expert identification. Green highlights ensuring correctness.

The prompt used for a GPT-4V-assisted visual instruction assessment are shown in Figure 16 and Figure 17.

GPT-4V-Assisted Prompt

{
"system_prompt" : "You are a helpful and precise assistant for checking the quality of the answer.",
"prompt_template" : "[Image Input] [Question]{question}
[Human] : {answer_0}[End of Human]
[Assistant 1] : {answer_1}[End of Assistant 1]
[Assistant 2] : {answer_2}[End of Assistant 2]
……
[Assistant 6] : {answer_6}[End of Assistant 6]",
"prompt" : "We request your evaluation of six AI assistants' responses to a user question based on an image. The user question is provided above and a reference image is uploaded. Additionally, a high-quality and accurate response from a human is given for comparison. Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.Please first output a single line containing only 6 values indicating the scores for Assistant 1 to 6, respectively. The six scores are separated by a space.In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment."
}

Figure 16: English prompt for guiding GPT-4V to rate the answers from GPT-4V, Gemini, LLaVA and MiniGPT-4.

GPT-4V-Assisted Prompt

{
"system_prompt" : "你是一位非常有帮助且精确的助手，用于检查答案的质量。",
"prompt_template" : "[Image Input] [问题]{question}
[人类] : {答案_0}[人类回复结束]
[助手1] : {答案_1}[助手1回复结束]
[助手2] : {答案_2}[助手2回复结束]
……
[助手6] : {答案_6}[助手6回复结束]",
"prompt" : "我们请求您根据一张图片，评估六位AI助手对用户问题的回答。用户的问题已经提供在上方，且参考图片已上传。此外，为了对比，还提供了一个人类给出的高质量且准确的回答。请评价它们的回答在有用性、相关性、准确性、细节层次上的表现。每位助手都会在1到10的范围内得到一个总分，分数越高表示总体表现越好。请首先输出仅包含6个分数的单行，分别代表助手1至6的得分，四个分数之间用空格分隔。在随后的行中，请提供一个全面的评价解释，避免任何潜在的偏见，并确保回答呈现的顺序不影响您的判断。"
}

Figure 17: Chinese prompt for guiding GPT-4V to rate the answers from GPT-4V, Gemini, LLaVA and MiniGPT-4.

OCR Prompt


```
"image": "ocr_image.jpg",
"conversations": [
  {
    "from": "human",
    "value": "<image>\n请识别图片中的文字"
  },
  {
    "from": "gpt",
    "value": "图片中显示的文字是：\"上海公司\""
  }
```

Figure 18: Chinese OCR prompt.

A.4 Data

Figure 19 presents an example of a knowledge graph with English content, where the instructions and responses were generated by GPT-4V following the specified English prompt.

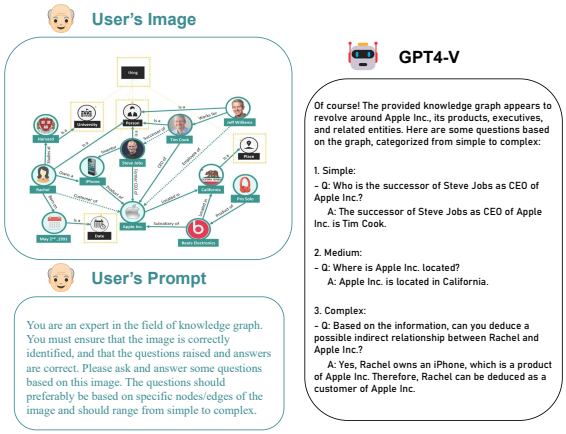


Figure 19: The instructions and answers followed by the user prompt for a given image in English.

Similarly, Figure 20 illustrates a flowchart with Chinese content, with instructions and responses produced by GPT-4V in accordance with the adapted Chinese prompt.

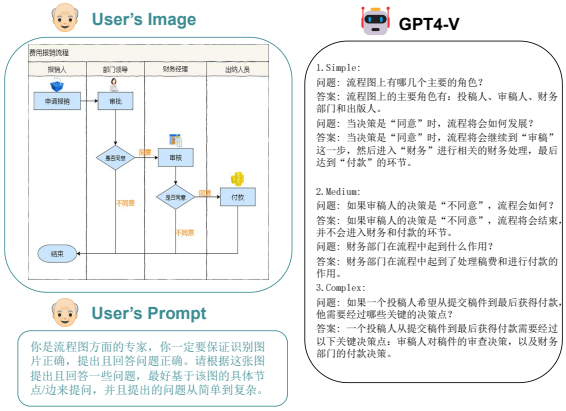


Figure 20: The instructions and answers followed by the user prompt for a given image in Chinese.

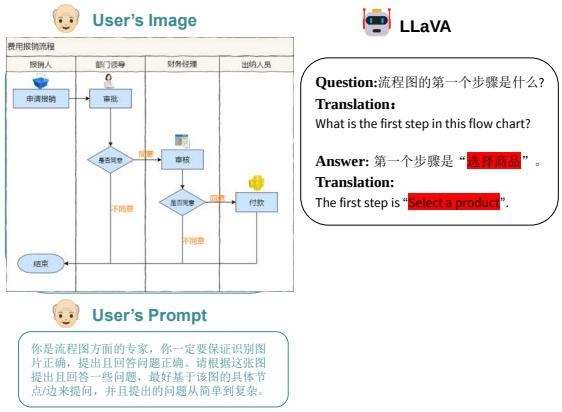


Figure 21: This figure is an illustration of the models' hallucination when responsing to Chinese instructions. Red highligths the hallucination in the response of the model. The meaning of the Chinese characters in the first step is "Apply for a refund" in Chinese while LLaVA recognized it as "Select a product" in Chinese.