FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis

Ziao Wang, **Yuhang Li**, **Junda Wu**, **Jaehyeon Soon**, **Xiaofeng Zhang*** ,

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
wangziao1993@hotmail.com, 331137797@qq.com, wujunda@stu.hit.edu.cn,
jaehyeon_soon@stu.hit.edu.cn, zhangxiaofeng@hit.edu.cn

Abstract

In this paper, we propose FinVis-GPT, a novel multimodal large language model (LLM) specifically designed for financial chart analysis. By leveraging the power of LLMs and incorporating instruction tuning and multimodal capabilities, FinVis-GPT is capable of interpreting financial charts and providing valuable analysis. To train FinVis-GPT, a financial task oriented dataset was generated for pre-training alignment and instruction tuning, comprising various types of financial charts and their corresponding descriptions. We evaluate the model performance via several case studies due to the time limit, and the promising results demonstrated that FinVis-GPT is superior in various financial chart related tasks, including generating descriptions, answering questions and predicting future market trends, surpassing existing state-of-the-art multimodal LLMs. The proposed FinVis-GPT serves as a pioneering effort in utilizing multimodal LLMs in the finance domain and our generated dataset will be release for public use in the near future to speedup related research.

1 Introduction

In the era of large language model (LLM) [6, 7, 9, 10], various real-world applications will be deeply and permanently changed by the LLMs as well as other large models (LMs). For instance, the LLMs already demonstrated a superior performance in various NLP tasks such as understanding and generating human-like text. Similarly, the large multimodal models (LMMs) has opened up new possibilities for more complex applications such as embodied robot. Thus, a good number of research efforts as well as industrial attentions have been attracted to explore the possibility whether such LMs could be utilized for financial related tasks.

Therefore, we are motivated to propose this novel multimodal large language model (FinVis-GPT) specifically designed for understanding financial chart. The proposed approach are two-stage ones. At the first stage, we must carefully prepare a dataset for this task which will be released for

---

*Corresponding author. Email: zhangxiaofeng@hit.edu.cn

public use in the near future. At the second stage, we train a large multimodal model using this dataset. Note that it is very demanding to tune a large multimodal model from the begining. Thus, we only fine-tune an existing model using this generated dataset. We expect that, by leveraging the power of LLMs, the proposed FinVis-GPT should be capable of interpreting financial charts and providing more accurate analysis in a human-like manner. This capability allows FinVis-GPT to answer a wide range of questions, such as predicting future trends based on historical data, identifying key patterns, and providing explanations for observed market phenomena.

As aforementioned, the key contribution of our work is the creation of a financial task oriented dataset for pre-training and instruction-tuning the large models. For the pre-training phase, we have curated a dataset comprising various types of financial charts along with their corresponding descriptions. This dataset enables FinVis-GPT to learn the intricate relationships between visual patterns in financial charts and their textual interpretations. For the instruction tuning phase, we have prepared a dataset that pairs images of financial charts with a set of instructions or questions. This dataset allows FinVis-GPT to learn how to respond to specific queries related to financial chart analysis, thereby enhancing its ability to generate relevant and accurate responses. After training the FinVis-GPT on this dataset, we investigate the model performance via various case studies due to the time limit. The results demonstrated that FinVis-GPT can effectively analyze financial charts and generate reliable and accurate interpretations. We believe that our work paves the way for more sophisticated applications of multimodal LLMs in the financial domain, potentially transforming how financial analysis is conducted.

2 Related Work

The evolution of LLMs and LMMs have already become the major research subjects recently. In this section, we briefly review several most pertinent works in these areas and discuss their relationship to our proposed model, FinVis-GPT.

Large Language Models and Instruction Tuning The transformation of LLMs into instruction followers has been a prominent research direction. For instance, InstructGPT [8] was introduced as a model designed to follow instructions given in natural language and generate useful responses.
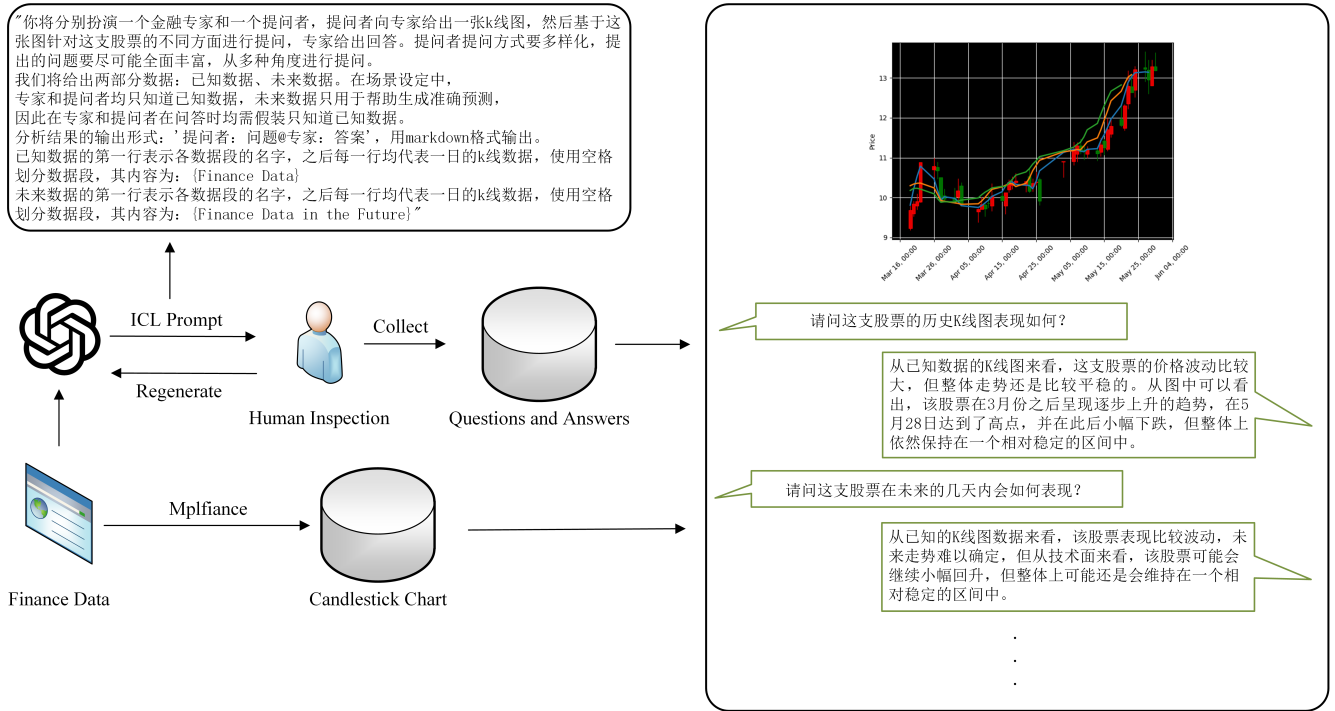
Figure 1: The designed process to generate multimodal dataset.

This model demonstrated that instruction tuning could significantly enhance the performance of LLMs, surpassing even the capabilities of GPT-3. Building on this concept, Chiang et al. [1] fine-tuned the LLaMA model [10] on user-shared dialogues collected from ShareGPT, resulting in an open-source chatbot with impressive performance.

Large Multimodal Models The extension of LLMs to handle multimodal inputs has been a significant advancement in recent research. The KOSMOS-1 model [4], trained from scratch on web-scale multimodal corpora, showcased impressive performance across language understanding, generation, and perception-language tasks. Similarly, MiniGPT-4 [12] demonstrated the potential of aligning a frozen visual encoder with a frozen LLM, Vicuna, using a single projection layer. Further extending the multimodal capabilities, mPLUG-Owl [11] was proposed to concurrently support multiple modalities and facilitate diverse unimodal and multimodal abilities through modality collaboration. In a similar vein, LLaMA-Adapter V2 [3] was proposed as a parameter-efficient model capable of handling visual instructions. Lastly, InstructBLIP [2] was designed to handle a variety of instructions, showcasing its ability to generate detailed captions, count specific objects, and address general inquiries posed by users.

Building upon these advancements, our proposed model, FinVis-GPT, incorporates financial charts as part of the multimodal input. This integration enables a more nuanced understanding of financial data, marking a significant step towards the application of multimodal LLMs in the financial domain. By leveraging the strengths of both instruction tuning and multimodal capabilities, FinVis-GPT aims to provide insightful analysis of financial charts, demonstrating the po-

tential of multimodal LLMs in domain-specific applications.

3 Generating Multimodal Financial Dataset

The data collection for FinVis-GPT involved creating datasets for two phases: pre-training alignment and instruction tuning. The goal of these datasets was to equip the model with the ability to understand and interpret multimodal data, particularly financial charts, and generate valuable responses based on given instructions. An illustrative example of our whole collection pipeline and the collected data is shown in Figure 1.

3.1 Pre-training Alignment Dataset

Pre-training alignment is a crucial step in training multimodal models, as it allows the model to align various types of data into a common embedding space. For the purpose of this step, we used historical daily stock price data of Chinese A-share from 2006 to 2023. This data was segmented into smaller sets containing 60-80 trading days, and each set was further divided into prompt data (data given to the model for prediction) and predict data (data to be predicted), with the former comprising 60-80% of each set.

Images were generated from this prompt data using the mplfinance[1] library, with a split of 80% for candlestick charts and 20% for line charts. To simulate real world scenarios, the generated charts were enhanced with moving averages of 3, 6, and 9 days, volume bars, and various chart styles, all added randomly.

---

[1]https://github.com/matplotlib/mplfinance

You will play the role of a financial expert. Upon receiving a k-line chart, you should first identify what type of chart it is and then describe the different stages of stock trends. You are required to conduct professional financial analysis on the input data while ensuring that your analysis is comprehensive and professional from different perspectives. Finally, you need to summarize your findings. To facilitate generating answers, you will not receive an image but rather data related to the k-line chart. In this scenario, since it is assumed that you are analyzing an image as an expert, your answer should pretend that you are analyzing an image and only mention content commonly found in k-line charts.

In your answer:

- Do not evaluate what you are doing; simply provide answers.

- Use "this stock" instead of direct stock codes.

- Do not explain the meaning of different data segments or their names.

- Do not draw charts; use text descriptions based on data only.

- Avoid saying more data is needed or suggesting other factors be considered; provide clear analytical conclusions instead.

The output format for analysis results: 'Answer', using markdown format. The first line of received content represents the name of each data segment, with each subsequent line representing one day's k-line data separated by spaces.

Table 1: Prompt designed for pre-training stage in data collection.

The data structure for each entry in this dataset consisted of an image, an instruction, and an answer. The instructions, designed to request an interpretation of the charts, were manually crafted. The answer for each instruction was generated by using chatGPT to interpret the prompt data. The prompt given to chatGPT are shown in Table 1.

3.2 Instruction Tuning Dataset

For instruction tuning, a separate dataset was created, comprising 200K sets, each with about five questions. The primary purpose of this dataset was to fine-tune FinVis-GPT's ability to generate relevant and accurate responses to specific queries related to financial chart analysis.

Like the pre-training alignment dataset, the data structure for this dataset also consists of an image, an instruction, and an answer. However, the key difference lies in the generation manner of instructions and answers: they were crafted using both prompt data and predict data, which allowed the model to generate future trend predictions, the prompt designed to give chatGPT is reported in Table 2.

3.3 Dataset Statistics

Table 3 provides a detailed breakdown of the key statistics associated with the collected datasets used in the pre-training and instruction tuning phases of the FinVis-GPT model. The count of words in questions, answers, and total dialog ex-

You will play the roles of a financial expert and a questioner respectively. The questioner presents a candlestick chart to the expert, and then asks questions based on different aspects of this stock. A total of five 'Q&A' sessions are required. The questioning should be diverse, with questions as comprehensive and rich as possible, from multiple perspectives. For example, asking the expert for professional financial analysis of the stock, predicting its future trend or explaining data. To facilitate prediction, we will provide two sets of data: known data and future data. In this scenario, it is assumed that both the expert and questioner only know about the known data; while future data is only used to help generate accurate predictions. Therefore, when answering questions related to predictions, experts must pretend that they only know about known data without mentioning future data explicitly but must combine them with generated predictions using future data to answer such questions clearly.

In addition, since you are playing an expert who receives images in this scenario; your answers need to use language suitable for analyzing an image-oriented perspective while focusing on content commonly found in candlestick charts. During questioning by the questioner or answering by the expert:

- Do not directly use stock codes; instead use 'this stock' as a substitute.

- Do not explain what each segment represents.

- Do not mention any segment names.

- Experts should refrain from evaluating their own tasks during responses.

- Avoid saying more information is needed or suggesting other factors need consideration.

- Do not allow questioner to view additional information themselves.

When asked for predictions by questioner during questioning sessions - experts must give out forecasts based on generated results using future-data. The output format for analytical results: "Question@Answer@" where both "@" symbols serve as separators between Question and Answer pairs except at any other time within either part (use markdown formatting). The first line of the known data indicates each segment's name followed by k-line trading records for each day, with data segments separated by spaces:
$\{prompt\_data\}$
The first line of the future data indicates each segment's name followed by k-line trading records for each day, with data segments separated by spaces:
$\{predict\_data\}$

Table 2: Prompt designed for instruction-tuning stage in data collection.

		mean	q-5%	q-95%
pre-train	# Question	28.68	21	36
	# Answer	401.15	179	882
	# Dialog	429.83	207	910
instruction	# Turns	4.79	3.00	5.00
	# Question	19.96	11.00	14.00
	# Answer	63.03	23	41
	# Dialog	397.36	238	748

Table 3: Summary statistics of the collected dataset. Here, '#' represents word count. The dataset includes two main categories: pre-training and instruction. The statistics cover the mean, 5th percentile (q-5%) and 95th percentile (q-95%) of word count for questions, answers, and dialogues in each category.
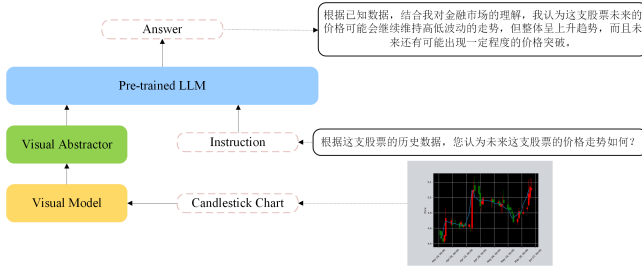

Figure 2: The model architecture.

changes (denoted as '#') are examined under various statistical metrics such as mean, the 5-th percentile (q-5%), and the 95th percentile (q-95%).

During pre-training, we observe that on average, questions have around 28.68 words, while answers contain approximately 401.15 words. This indicates that responses tend to be much more detailed and comprehensive. The entire dialog, including both questions and answers, contains about 429.83 words on average. The data distributions for the number of words in the questions, answers, and the entire dialog show a wide spread, as evidenced by the 5th and 95th percentile values.

In the instruction tuning phase, the number of turns taken averages at 4.79, hinting at the complexity and depth of the conversations in the dataset. Questions contain fewer words compared to the pre-training dataset, with an average of 19.96 words. The answers in this phase are significantly shorter, with approximately 63.03 words on average. This suggests a shift towards more focused and concise communication. The entire dialog contains about 397.36 words on average, with a less pronounced spread than observed in the pre-training dataset.

4 Model Training

The FinVis-GPT model was built on top of the pre-existing LLaVA [5] model, incorporating the advanced language capabilities of the latter while extending them for the specific financial context, the model architecture is plotted in Figure 2. The training process consists of two major steps: pre-training alignment and instruction tuning.

4.1 Pre-training Alignment

Pre-training alignment aimed at teaching the model to understand the relationship between visual patterns in financial charts and their corresponding textual interpretations. The pre-training alignment dataset, consisting of various financial charts and corresponding descriptions, was used for this purpose.

For the pre-training, we adopted the same training approach as LLaVA but used our specifically curated dataset of financial charts and descriptions. The model was trained using a batch size of 128 and a learning rate of 2e-3. The pre-training was carried out on 8 NVIDIA Tesla A100 GPUs for 1 epochs.

The effectiveness of pre-training alignment was evaluated by feeding the model with new, unseen financial charts and checking its ability to generate accurate and relevant descriptions. The generated descriptions were evaluated by a panel of financial experts for their accuracy and relevance.

4.2 Instruction Tuning

Instruction tuning is a technique that allows the model to learn how to generate appropriate responses to specific instructions or queries. For this, we used the instruction tuning dataset, which was specifically created for the purpose of fine-tuning FinVis-GPT.

The tuning phase involved adjusting the model's parameters so that it could accurately respond to instructions about financial charts. This phase was also executed using a batch size of 128 and a learning rate of 1e-5 for 3 epochs.

4.3 Regularization and Model Validation

To prevent overfitting during the training process, we incorporated dropout and weight decay regularization techniques. We also used early stopping based on the validation set performance to determine the optimal number of training epochs.

Model validation was performed intermittently throughout the training process. We maintained a holdout validation set that was not used during the training process. At the end of each epoch, the model was tested on this validation set to gauge its performance and to ensure it was learning the intended tasks effectively.

In sum, the training process of FinVis-GPT was a meticulous process aimed at harnessing the language prowess of LLaVA and tailoring it to the complex task of financial chart interpretation and analysis.

5 Experiments

5.1 Experimental Setup

We compared FinVis-GPT against several baseline models including LLaVA [5], MPLUG-Owl [11], and MiniGPT-4 [12]. Each of these models represents the latest advancements in multimodal learning with unique advantages. The metrics used for comparison included quality of financial chart descriptions, understanding of financial context, and prediction accuracy of financial trends. We employed the following three tasks to evaluate each model:
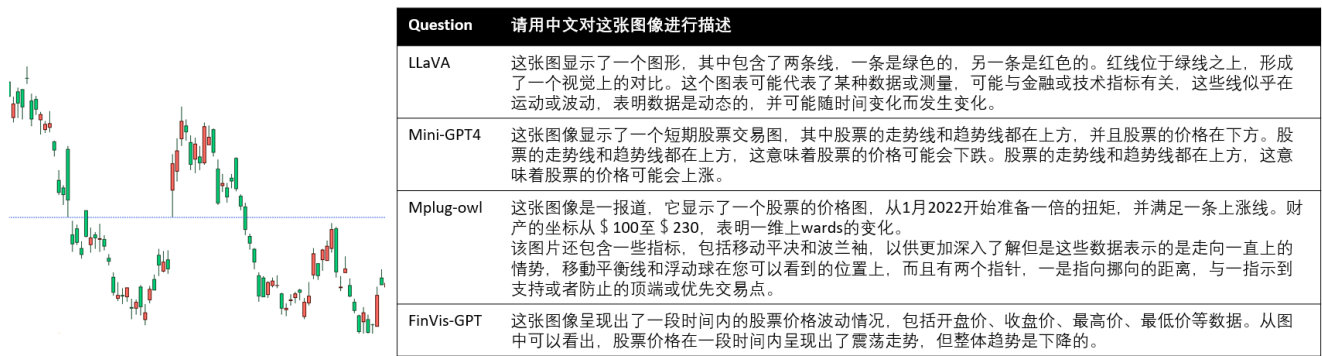
Question	请用中文对这张图像进行描述
LLaVA	这张图显示了一个图形，其中包含了两条线，一条是绿色的，另一条是红色的。红线位于绿线之上，形成了一个视觉上的对比。这个图表可能代表了某种数据或测量，可能和金融或技术指标有关，这些线似乎在运动或波动，表明数据是动态的，并可能随时间变化而发生变化。
Mini-GPT4	这张图像显示了一个短期股票交易图，其中股票的走势线和趋势线都在上方，并且股票的价格在下方。股票的走势线和趋势线都在上方，这意味着股票的价格可能会下跌。股票的走势线和趋势线都在上方，这意味着股票的价格可能会上涨。
Mplug-owl	这张图像是一报道，它显示了一个股票的价格图，从1月2022开始准备一倍的扭矩，并满足一条上涨线。财产的坐标从＄100至＄230，表明一维上wards的变化。 该图片还包含一些指标，包括移动平决和波兰袖，以供更加深入了解但是这些数据表示的是走向一直上的情势，移动平衡线和浮动球在您可以看到的位置上，而且有两个指针，一是指向挪向的距离，与一指示到支持或者防止的顶端或优先交易点。
FinVis-GPT	这张图像呈现出了一段时间内的股票价格波动情况，包括开盘价、收盘价、最高价、最低价等数据。从图中可以看出，股票价格在一段时间内呈现出了震荡走势，但整体趋势是下降的。

Figure 3: Experiment results on description generation.

Question	请解释这张图像中的黑线和白线表达什么意思
LLaVA	这张图像中的黑线和白线是用来表示不同的数据或信息。在这种情况下，黑线代表了过去的数据或历史数据，而白线则表示了预测或估计的数据，也就是未来的趋势。这种图表的设计帮助人们更直观地了解投资品种的表现，并估算未来的可能趋势。
Mini-GPT4	这张图中的黑线表示股票价格的上涨趋势，而白线表示股票价格的下跌趋势。May 2023 根据这张图，股票价格在2020年5月10日至2020年12月31日期间上涨，但在2021年1月1日至2021年1月15日期间下跌。此外，股票价格在2020年6月1日至2020年8月31日期间上涨，但在2020年9月1日至2020年10月31日期间下跌。因此，可以看出股票价格在2020年6月1日至2020年8月31日期间上涨，但在2020年9月1日至2020年10月31日期间下跌。
Mplug-owl	图中显示一条黑线（日期横线）和一折线的叠合表示日志数据的变化。银行日程架构为一直滚动的线条，而执行者日记数量为整数，每日的总数和排列数组的最大值。拓展概感，某一日在泊线插入一点数值，表明档捷的积分。一般，增加植树材料的数目，可能会引起历史数学梯度的改变。
FinVis-GPT	在这张k线图中，黑线代表收盘价低于开盘价，表示当天股票价格下跌；白线代表收盘价高于开盘价，表示当天股票价格上涨。这是一种常用的方法来展示股票价格走势。

Figure 4: Experiment results on question answering.

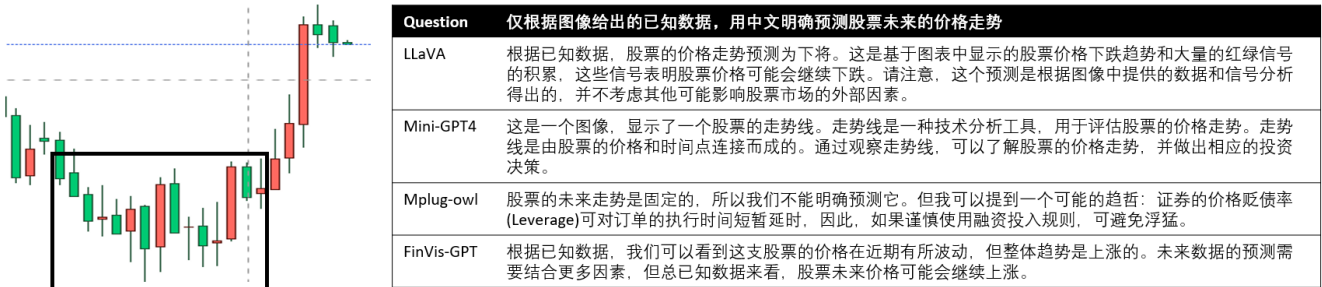| Question	仅根据图像给出的已知数据，用中文明确预测股票未来的价格走势
LLaVA	根据已知数据，股票的价格走势预测为下将。这是基于图表中显示的股票价格下跌趋势和大量的红绿信号的积累，这些信号表明股票价格可能会继续下跌。请注意，这个预测是根据图像中提供的数据和信号分析得出的，并不考虑其他可能影响股票市场的外部因素。
Mini-GPT4	这是一个图像，显示了一个股票的走势线。走势线是一种技术分析工具，用于评估股票的价格走势。走势线是由股票的价格和时间点连接而成的。通过观察走势线，可以了解股票的价格走势，并做出相应的投资决策。
Mplug-owl	股票的未来走势是固定的，所以我们不能明确预测它。但我可以提到一个可能的趋哲：证券的价格贬债率(Leverage)可对订单的执行时间短暂延时，因此，如果谨慎使用融资投入规则，可避免浮猛。
FinVis-GPT	根据已知数据，我们可以看到这支股票的价格在近期有所波动，但整体趋势是上涨的。未来数据的预测需要结合更多因素，但总已知数据来看，股票未来价格可能会继续上涨。

Figure 5: Experiment results on trend prediction.

- **Description Generation:** For this task, the models were given an image of a financial chart and were required to generate a description, capturing the key trends, patterns, and anomalies.

- **Question Answering:** This task involved a comprehension test where models were given an image of a financial chart along with a set of questions. The questions were designed to assess the model's understanding of the financial context of the chart.

- **Trend Prediction:** For this task, models were provided an image of a financial chart along with historical financial data and were asked to predict future trends. The predictions were compared with actual future data to evaluate the model's predictive performance.

5.2 Results and Discussion

Description Generation. The task of description generation is exemplified in Figure 3, where a randomly selected outcome is presented. Based on these results, it is obvious that LLaVA fails to accurately identify the image as a representation of stock trends. In contrast, Minit-GPT4 demonstrated a superior understanding by correctly recognizing the image as a stock trading chart, though it inaccurately identified the blue line as a stock trend line. Moreover, mplug-owl managed to acknowledge the image as a stock price chart but it introduced several unrelated elements, causing its description to veer off the accurate interpretation. Among all models assessed, FinVis-GPT emerged as the most proficient, correctly recognizing the image and providing a concise and accurate description. This underscores its capacity for generating superior descriptions when compared to the other models in this specific context.

Question Answering. The question answering task is plotted in Figure 4. The results reveal that LLaVA substantially misconstrued the stock trend, erroneously identifying the black candle line as the past trend and the white as the future trend. Meanwhile, Mini-GPT4 muddled the representation of black and white lines, further compounding its output with a significant amount of irrelevant content. The mplug-owl model exhibited a complete lack of recognition for the image, fabricating an entirely unrelated narrative. In contrast, the response provided by FinVis-GPT was both concise and accurate, earning it the top spot amongst the compared models for this task. Its output underscores the superior efficacy of FinVis-GPT in understanding and accurately answering questions based on the given visual representation.

Trend Prediction. An example of trend prediction is depicted in Figure 5. The left image represents a market trend over a certain period, with the trend within the black box provided as input to the models. The accurate prediction for this trend should indicate an upward trajectory. However, LLaVA's prediction was contrary to this, presenting a downward trend instead. Mini-GPT4 failed to answer the prediction question accurately, and instead produced unrelated information, a phenomenon often referred to as 'hallucination'. Similarly, mplug-owl's output was also characterized by this 'hallucinating' issue. In contrast, FinVis-GPT's prediction

was not only accurate but also incorporated a proper description of the trend. This showcases FinVis-GPT's superiority in trend prediction tasks, with an ability to provide both accurate and informative responses.

6 Conclusion

In this work, we presented FinVis-GPT, a novel large multi-modal model tailored to the financial domain, with a focus on financial chart analysis. Our approach integrated the benefits of pre-trained LLMs with a curated dataset sourced directly from the financial sector. The FinVis-GPT model showed significant improvement over existing models in terms of generating accurate, relevant, and financially styled responses. Through the creation of a robust instruction tuning dataset and case studies, we have demonstrated the potential of multi-modal LLMs in the financial sector. This work lays the foundation for more sophisticated applications of AI in finance, potentially transforming the landscape of financial analysis. Future work will focus on further expanding the applicability of FinVis-GPT in more diverse financial scenarios and real-time financial decision-making.

References

[1] 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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[2] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023. URL http://arxiv.org/abs/2305.06500.

[3] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model, 2023. URL http://arxiv.org/abs/2304.15010.

[4] 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, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language Is Not All You Need: Aligning Perception with Language Models, 2023. URL http://arxiv.org/abs/2302.14045.

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, 2023. URL http://arxiv.org/abs/2304.08485.

[6] OpenAI. Introducing ChatGPT, 2022. URL https://openai.com/blog/chatgpt.

[7] OpenAI. Gpt-4 technical report, 2023.

[8] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

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

[10] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, 2023. URL http://arxiv.org/abs/2303.03846.

[11] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, 2023. URL http://arxiv.org/abs/2304.14178.

[12] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, 2023. URL http://arxiv.org/abs/2304.10592.