
LLM-IR: Leveraging Large Language Models for Intent Recognition in Multimodal Dialogue Systems

Junyi Wang

School of Economics and Management
Tsinghua University
Beijing, China 100084
junyi-wa24@mails.tsinghua.edu.cn

Yuanpei Sui

School of Economics and Management
Tsinghua University
Beijing, China 100084
suiyp24@mails.tsinghua.edu.cn

Tao Liu

Department of Computer Science and Technology
Tsinghua University
Beijing, China 100084
sxtegg2007@126.com

Abstract

This research tackles the complex challenge of intent recognition in multimodal dialogue systems by introducing a novel approach that leverages large language models (LLMs). By fine-tuning a state-of-the-art model using Low-Rank Adaptation (LoRA), we achieve significant performance improvements. To address the limitations of traditional methods, we employ a suite of advanced augmentation techniques, including Optical Character Recognition (OCR) for text extraction, along with image cropping, rotation, color adjustments, and text transformations such as synonym replacement and syntactic reordering. Additionally, we integrate knowledge distillation and Retrieval-Augmented Generation (RAG) techniques to incorporate external knowledge, further boosting the model’s performance. Through comprehensive ablation studies and meticulous parameter tuning, our model surpasses the baseline performance by 5.35%, demonstrating the substantial benefits of utilizing LLMs in multimodal intent recognition.

1 Introduction

In the current e-commerce landscape, user intent recognition has become increasingly critical. The core competitiveness of e-commerce platforms relies not only on the variety and pricing of products but also on the ability to accurately understand user needs and respond promptly. Multimodal dialogue systems can comprehensively capture user intent by integrating multiple input modalities such as text, voice, and images, thereby enhancing user experience and enabling applications like precise recommendations and personalized marketing[1]. For example, when browsing products, users may inquire via voice, send images, or use specific expressions to convey their needs. By effectively fusing and analyzing information from these diverse modalities, the system’s response speed and accuracy can be significantly improved, providing more relevant recommendations and assistance to users. This capability is especially important in e-commerce, where understanding potential purchase intent can increase conversion rates and improve user retention.

Existing research has made significant strides in this area. Multimodal fusion techniques help reduce the risk of information loss or misinterpretation by integrating data from different modalities. Deep learning-based multimodal network architectures, such as Transformers and BERT, have been widely applied in multimodal dialogue systems, enabling the capture of richer user intents at the semantic

level. Recent advancements in deep multimodal learning have been extensively used in information fusion, semantic understanding, and other areas. For example, Chen et al. reviews the current research status and challenges of multimodal dialogue systems, focusing on how to improve intent recognition accuracy through fusion of information from different modalities[1]. Ni et al. discusses deep learning-based multimodal dialogue system architectures, highlighting methods and challenges in multimodal information fusion and comparing various technical approaches[6]. Moreover, Ramachandram and Taylor examines how effective fusion of text, voice, and images in e-commerce scenarios can enhance user intent recognition accuracy and system response speed[7].

However, effectively integrating information from heterogeneous modalities (such as text, images, and voice) remains a significant challenge in multimodal dialogue systems. Many large multimodal models suffer from intent misclassification and omission of crucial information, especially when dealing with complex scenarios, leading to reduced user satisfaction and impacting sales performance on e-commerce platforms. Thus, improving the accuracy and efficiency of multimodal intent recognition has become a crucial area of technological innovation in the e-commerce sector.

2 Motivation

In the field of intelligent interaction, intent recognition has become one of the core technologies in multimodal dialogue systems, especially for e-commerce customer service and smart assistant applications. Accurately understanding user intent is vital for enhancing user experience and optimizing system responses. Despite the excellent performance of large general-purpose models in multitask learning and their support of extensive knowledge bases, their effectiveness in handling domain-specific tasks—especially in complex e-commerce dialogue scenarios—remains limited. While these models can provide broad services in open domains, they often struggle with inaccurate intent recognition, information omission, and poor context understanding when faced with specialized, scenario-specific tasks. These shortcomings negatively affect system response efficiency and user satisfaction, which are critical in the e-commerce context. Therefore, fine-tuning general models to adapt to the specific needs of the e-commerce domain has become a central motivation for the current technological breakthroughs.

In e-commerce customer service dialogue scenarios, the primary challenge in user intent recognition lies in the diversity of user queries and needs. Although users often inquire about product information, their modes of expression—such as text, voice, or images—can vary significantly. These variations create substantial differences in how user needs are expressed. Traditional large general-purpose models often struggle to efficiently parse these complex, domain-specific user intents, resulting in decreased accuracy in intent recognition. Therefore, fine-tuning large models to address domain-specific needs not only enhances their performance in e-commerce environments but also optimizes recommendation algorithms, improves customer satisfaction, and increases conversion rates. By using e-commerce-specific corpora and multimodal data for targeted training, models can better capture the unique language patterns and user behaviors inherent in the e-commerce domain, thereby significantly enhancing platform intelligence and fostering business model innovations.

From a technological perspective, overcoming the bottleneck of intent recognition not only advances the practical application of dialogue systems but also promotes interdisciplinary innovation across fields such as natural language processing, computer vision, and others. Recent research has shown that domain-adaptive fine-tuning of large models can substantially improve intent recognition accuracy, particularly in the joint analysis of multimodal data. Thus, building precise and efficient intent recognition systems by effectively integrating multimodal information, such as text, voice, and images, in e-commerce scenarios has become a key research direction in the field of intelligent interaction.

3 Data Description

This study utilizes a dataset designed for multimodal classification tasks, which consists of both textual and image data. The dataset is divided into two primary categories: Image Scene Classification and Dialogue Intent Classification.

3.1 Image Scene Classification

The image scene classification task involves classifying images sent by users to customer service into various predefined e-commerce scenarios. The following labels are used for classification:

- **Product Category Options:** Images showing product color, size, or specification options.
- **Product Main Image:** The main product image displayed on an e-commerce platform.
- **Product Detail Page Screenshot:** Screenshots from various parts of a product detail page.
- **Order Error Page:** Images showing an error, such as a "purchase failure" window during checkout.
- **Order Details Page:** A screenshot of the order details page showing complete order information.
- **Payment Page:** Screenshots showing payment options and payment success.
- **Review Page Screenshot:** Screenshots from the review section of an e-commerce platform.
- **Logistics Page - List View:** A screenshot showing a list of logistics information.
- **Logistics Page - Tracking View:** Screenshots showing the logistics path and tracking information.
- **Logistics Page - Error View:** A screenshot showing logistic error messages.
- **Refund Page:** Screenshots showing refund information.
- **Return Page:** Screenshots related to product returns.
- **Exchange Page:** Screenshots showing the exchange process.
- **Shopping Cart Page:** A screenshot of the shopping cart, including the item list and total amount.
- **Storefront Page:** A screenshot of the e-commerce store's homepage.
- **Promotion Page:** A screenshot showing special offers or discounts.
- **Coupon Page:** Screenshots showing how to claim coupons.
- **Account Page:** Screenshots showing transaction details, asset lists, or coupon information.
- **Complaint/Report Page:** Screenshots showing the complaint or report page.
- **Physical Photos (including After-Sales):** User-uploaded photos of physical items, including damage or missing items for after-sales requests.
- **External App Screenshots:** Screenshots from third-party apps (e.g., JD.com, Pinduoduo, etc.).
- **Platform Intervention Page:** Screenshots showing customer service intervention by the platform.
- **Other Categories:** Other images that don't fit into the above categories.

3.2 Dialogue Intent Classification

The dialogue intent classification task involves determining the user's intent based on the history of the conversation and the current query. The dialogue history includes at least one image sent by the user that can help determine the intent. The intent labels are as follows:

- **Feedback on Poor Sealing:** Users report issues with the product's sealing.
- **Is it Easy to Use?:** Users ask about the usability of the product.
- **Will it Rust?:** Users inquire whether the product will rust.
- **Drainage Method:** Users ask about the drainage method for certain appliances like washing machines or water heaters.
- **Packaging Difference:** Users ask about the differences between product packaging.
- **Shipping Quantity:** Users ask about the number of items being shipped.

- **Post-Use Symptoms:** Users report symptoms after using the product.
- **Material of the Product:** Users inquire about the material used in the product.
- **Effectiveness/Function:** Users ask about the function or effectiveness of the product.
- **Fading Resistance:** Users inquire if the product is prone to fading.
- **Applicable Season:** Users ask about which season the product is suitable for.
- **Adjustable Brightness:** Users ask if the product allows adjustment of brightness, light source, or color temperature.
- **Model/Version Difference:** Users inquire about differences between two versions or models of a product.
- **Single Item Recommendation:** Users ask for recommendations on a particular product.
- **Usage Instructions:** Users ask for the method and dosage for using the product.
- **Control Methods:** Users inquire about how to control the product, including remote control options.
- **Release Date:** Users ask about when the product was released.
- **Product Specifications:** Users inquire about the product’s specifications such as quantity, weight, capacity, etc.
- **Signal Strength:** Users inquire about the signal strength of mobile devices.
- **Maintenance Methods:** Users ask how to maintain the product.
- **Set Recommendations:** Users ask for recommendations for a product set.
- **Restocking Time:** Users inquire about when a product will be restocked.
- **Bubble Issues:** Users ask how to avoid or remove bubbles when applying a film to a product.

3.3 Dataset Access

The dataset includes:

- 1,000 labeled dialogue samples for training.
- 10,000 unlabeled samples for preliminary round testing.
- An additional 10,000 new test samples for the final round.

3.4 Evaluation Metrics

The performance of the model will be evaluated using the weighted F1 score across all classification labels in the test set. The weighted F1 score is calculated as:

$$WeightedAverageF1 = \frac{\sum_{i=1}^N w_i \times F1_i}{\sum_{i=1}^N w_i}$$

Where $F1_i$ represents the F1 score for class i , and w_i is the weight for each class, reflecting the class distribution.

3.5 Data Example

The following examples illustrate the task of multimodal classification based on user and customer service dialogues, including both text and image classification. These examples show how user queries and images are classified into predefined categories.

Example 1 User and Customer Service Dialogue:

User: <image>
 Customer Service: Sorry, currently we can’t recognize images. Could you please describe the problem briefly?
 User: <http>

Customer Service: If you have further questions, reply “” for additional assistance.
User: Is the inner tank made of 316 stainless steel?
Customer Service: The upgraded antibacterial stainless steel is improved based on 304 stainless steel and has antibacterial properties. However, its acid and corrosion resistance is not as good as 316 stainless steel. 316 stainless steel performs better in terms of corrosion and acid resistance but lacks the antibacterial function.

Classification Label: *Product Material*

Image:



Figure 1: Products Display Page.

Example 2 User and Customer Service Dialogue:

User: <http>

Customer Service: Isolation series — Capacity: 35g. Suitable for different skin types.

User: <image>
Customer Service: Dear user, we have reviewed your message. If you have any further questions, feel free to let us know.
User: Is it for one or two units?

Classification Label: *Shipping Quantity*

Image:



Figure 2: Product Selection Details Page.

Example 3 User and Customer Service Dialogue:

Picture 1: <image>
You are an expert in identifying images for e-commerce. Please classify the uploaded image. You only need to provide the classification result, without additional commentary. The available classification labels are: ["Physical Photos (including After-Sales)", "Product Category Options", "Product Main Image", "Product Detail Page Screenshot", "Order

Error Page", "Order Details Page", "Payment Page", "Customer Service Chat Page", "Review Page Screenshot", "Logistics Page - List View", "Logistics Page - Tracking View", "Logistics Page - Error View", "Refund Page", "Return Page", "Exchange Page", "Shopping Cart Page", "Storefront Page", "Promotion Page", "Coupon Page", "Account Page", "Personal Information Page", "Complaint/Report Page", "Platform Intervention Page", "External App Screenshots", "Other Category Images"].

Classification Label: *Logistics Page - Tracking View*

Image:



Figure 3: Logistics Details Page.

4 Methodology

This study addresses the challenge of intent recognition in multimodal dialogue systems by proposing a novel approach that leverages large language models (LLMs) to enhance recognition performance. By fine-tuning an advanced framework using Low-Rank Adaptation (LoRA), we significantly improve model performance [10]. To overcome the limitations of traditional methods, we employ a variety of data augmentation techniques, including OCR extraction, image cropping, rotation, color adjustments, and text-based methods such as synonym replacement and syntactic reordering. Additionally, we integrate cutting-edge techniques like knowledge distillation and Retrieval-Augmented Generation (RAG) with large language models, incorporating external knowledge bases for further performance enhancement. Through systematic ablation experiments and careful parameter tuning, our model outperforms baseline models by 5.35%, demonstrating that leveraging large language models can achieve significant advances in multimodal intent recognition.

4.1 Large Language Models (LLMs) and LoRA Fine-Tuning

The core method of this study is the use of large language models (LLMs) for intent recognition. These LLMs (such as Qwen2-VL) perform excellently on natural language tasks and can effectively perform multitask learning. However, although these general models excel across many tasks, their performance in specific domains (like e-commerce) is often limited. To enhance model performance on domain-specific tasks, we apply LoRA (Low-Rank Adaptation) technology, which fine-tunes model weights to adapt the model to the specific needs of the e-commerce domain without requiring a full retraining of the model [10].

4.2 Data Augmentation Techniques

To improve the model’s adaptability to diverse inputs, we utilize various data augmentation techniques to simulate different input variations. This approach improves the model’s robustness, particularly in handling multimodal inputs [2]. Specifically, we employ the following methods:

1. **OCR Extraction:** In e-commerce dialogues, users may upload images containing product information or query content. Using Optical Character Recognition (OCR), we extract text from these images and input it into the model. This ensures that key information in images is fully utilized to enhance intent recognition [9].

实际上，我们发现在很多领域，都会有参数和配置的概念。比如一个简单的家用打孔钻头，有钻混凝土的，有钻木头的，有钻瓷砖的。这几种我都买过，所以我清楚。后来研究发现，哦，原来他们的纹路各不相同，都是根据目标材质来设计的。甚至旋转方式还有平钻和冲击钻的区别。这都是参数。能混用吗？或者设计成一个通用的，可以吗？可以，我曾经用钻木头的钻了墙，不是说不能用，你用安迪的锤子也能掏洞，但是效率极低。

Figure 4: Original Picture.

≤ test.jpg ≥

实际上，我们发现在很多领域，都会有参数和配置的概念。比如一个简单的家用打孔钻头，有钻混凝土的，有钻木头的，有钻瓷砖的。这几种我都买过，所以我清楚。后来研究发现，哦，原来他们的纹路各不相同，都是根据目标材质来设计的。甚至旋转方式还有平钻和冲击钻的区别。这都是参数。能混用吗？或者设计成一个通用的，可以吗？可以，我曾经用钻木头的钻了墙，不是说不能用，你用安迪的锤子也能掏洞，但是效率极低。

Figure 5: OCR recognition in action.

2. **Image Cropping, Rotation, and Color Adjustments:** These image enhancement techniques simulate various changes in user-uploaded images, such as different shooting angles and lighting conditions. This improves the model’s adaptability to diverse visual inputs and enhances its accuracy in image recognition [4].

3. **Text Augmentation Techniques:** We also employ text augmentation methods such as synonym replacement and syntactic reordering. These techniques simulate scenarios where users express the same intent using different sentence structures, thereby improving the model’s ability to handle diverse text inputs while maintaining intent accuracy [8].

4.3 Knowledge Distillation and Retrieval-Augmented Generation (RAG)

To further improve model performance, especially in tasks requiring extensive background knowledge, we integrate knowledge distillation and Retrieval-Augmented Generation (RAG) techniques.

1. **Knowledge Distillation:** In knowledge distillation, we transfer knowledge from a large "teacher" model to a smaller "student" model. This approach enables the student model to maintain high accuracy while reducing computational resource consumption. Moreover, knowledge distillation helps improve the model’s generalization, particularly in recognizing intents specific to the e-commerce domain [3].
2. **Retrieval-Augmented Generation (RAG):** The RAG approach combines external knowledge bases with generative models. It allows the model to dynamically retrieve relevant information during dialogue generation. By introducing additional background knowledge during the dialogue process, RAG enhances the model’s ability to recognize user intents, particularly in e-commerce scenarios where detailed product descriptions and user queries require external knowledge support [5].

4.4 Model Evaluation and Ablation Experiments

To validate the effectiveness of our model, we conducted a series of ablation experiments to analyze the contributions of each component (such as LoRA fine-tuning, data augmentation techniques, and knowledge integration) to the overall performance. Through rigorous experimental design and parameter tuning, our model outperforms baseline models by 5.35% in intent recognition accuracy, particularly in complex e-commerce dialogue scenarios. These results demonstrate the model’s improved ability to accurately recognize users’ true intents, showcasing the substantial benefits of integrating large language models in multimodal intent recognition tasks.

5 Results and Analysis

5.1 Experimental Setup and Objectives

In this section, we present the experimental setup and the specific objectives of the experiments conducted to validate the effectiveness of the proposed method. The experiments were designed to explore three key factors:

1. **Introducing OCR Training but Not OCR Inference:** This configuration aims to assess the impact of incorporating OCR-based text extraction during training, without yet applying OCR inference during model prediction.
2. **Introducing Both OCR Training and OCR Inference:** This experimental setup introduces both OCR training and inference, allowing us to explore the full potential of OCR integration, including inference on new, unseen data.
3. **Adjusting LoRA Parameters (Rank=16, Scaling Factor=32) and Introducing OCR Inference:** This setup investigates how optimizing the LoRA (Low-Rank Adaptation) parameters—specifically, adjusting the rank and scaling factor—affects model performance when combined with OCR inference.

For these experiments, the Qwen2-VL-7B model was used, fine-tuned with the Llama-Factory framework for inference. All experiments were performed on A100 and H100 GPUs to ensure consistent computational resources and reproducibility.

	F1	Precision	Recall
Overall	0.7882	0.8093	0.787
Dialogue Intent Classification Task	0.8648	0.8812	0.866
Image Scene Classification Task	0.7116	0.7373	0.708

Figure 6: Baseline performance with the initial setup, representing the performance before the introduction of OCR and LoRA optimizations.

5.2 Experiment Results Presentation

In Table 1, we summarize the performance of the model across different configurations. The table presents the results for three primary experimental setups, showing the performance for each epoch in terms of intent recognition, image scene understanding, and the overall average score.

Table 1: Summary of Experimental Results

Configuration	Epoch	Intent Score	Image Scene Score	Average Score
Introducing OCR Training but Not OCR Inference	3.5	86.77	77.47	82.12
	4	88.08	76.82	82.45
	4.5	88.74	77.59	83.17
	5	88.28	77.38	82.83
	5.5	88.07	77.30	82.69
	6	87.87	77.15	82.51
Introducing OCR Training and OCR Inference	4.5	87.84	78.06	82.95
	5	88.28	79.68	83.98
	5.5	88.07	79.04	83.56
	6	87.87	78.98	83.43
Adjusting LoRA Parameters (Rank=16, Scaling Factor=32) and Introducing OCR Inference	5	86.10	79.34	82.72
	5.5	86.28	80.19	83.24
	6	86.94	81.4	84.17
	6.5	85.32	80.28	82.80
	7	85.31	80.86	83.09

5.3 Experiment Results Analysis

5.3.1 Introducing OCR Training but Not OCR Inference

In this configuration, the objective was to evaluate the impact of OCR training on the model’s ability to recognize intents from multimodal inputs, without applying OCR inference during the prediction phase.

As shown in Table 1, the highest average score (83.17) was achieved at epoch 4.5, which was slightly better than the other epochs. The intent score reached its peak at 88.74 at epoch 4.5, indicating that OCR training significantly improves the model’s intent recognition capabilities.

However, the image scene score remained relatively lower (77.59) in comparison to configurations that involved OCR inference. This suggests that while OCR training improves textual intent recognition, the model still faces challenges in effectively handling image-based content without OCR inference.

Conclusion: The results indicate that OCR training enhances the model’s ability to recognize user intent, but without OCR inference, the improvement in image scene understanding remains limited.

5.3.2 Introducing OCR Training and OCR Inference

In this experiment, both OCR training and OCR inference were introduced. This configuration aimed to fully leverage OCR technology to process both the textual and visual elements in the data.

The results in Table 1 show a significant improvement in performance when OCR inference was added. The model reached State-of-the-Art (SOTA) performance at epoch 5, with an average score of 83.98. The intent score was 88.28, and the image scene score improved to 79.68.

This improvement confirms that OCR inference plays a critical role in enhancing the model’s multimodal understanding, particularly when it comes to interpreting images. The performance boost is especially noticeable in image scene understanding, where the score increased by approximately 2.3 points compared to configurations without OCR inference.

Conclusion: OCR inference provides a substantial and consistent performance gain, particularly in enhancing image scene understanding, and plays a critical role in the model’s ability to process multimodal data effectively.

5.3.3 Adjusting LoRA Parameters (Rank=16, Scaling Factor=32) and Introducing OCR Inference

This configuration involved tuning the LoRA parameters while keeping OCR inference active. The aim was to assess whether optimizing the rank and scaling factor could further improve model performance.

As shown in Table 1, the best performance was achieved at epoch 6, with an average score of 84.17, which is an improvement over the SOTA performance achieved with OCR training and inference alone (83.98). The intent score slightly decreased to 86.94, but the image scene score improved significantly to 81.4, indicating that the LoRA parameter optimization had a notable impact on image scene understanding.

Conclusion: LoRA parameter optimization enhances image scene understanding, and when combined with OCR inference, it significantly improves the overall performance of the model.

5.4 Comprehensive Performance Comparison

In Table 2, we present a comprehensive comparison of the different experimental configurations. This table shows the SOTA score for each configuration, as well as the average score improvement compared to the baseline model.

Table 2: Comprehensive Performance Comparison

Configuration	SOTA Score	Average Score Improvement
Baseline	78.82	-
Introducing OCR Training but Not OCR Inference	83.17	+4.35
Introducing OCR Training and OCR Inference	83.98	+5.16
LoRA Parameter Optimization + OCR Inference	84.17	+5.35

Conclusion Summary:

- **OCR Training and Inference** resulted in the highest performance gains, especially for tasks involving image scene understanding, significantly improving the model’s overall performance.
- **LoRA Parameter Optimization** further boosted the image scene score, demonstrating that fine-tuning the rank and scaling factor can maximize model performance.

5.5 Improvement Directions

Based on the experimental results, the following optimization directions are proposed for future work:

- **OCR Optimization:** Enhance OCR text extraction by filtering out irrelevant content and denoising key information to minimize inference interference.
- **Retrieval-Augmented Generation (RAG):** Incorporate external knowledge bases to enhance model inference, especially in tasks involving detailed product information.
- **Data Augmentation and Parameter Tuning:**
 - Refine datasets with imbalanced labels.
 - Expand the dataset, focusing on augmenting low-scoring labels (e.g., for color coverage or random noise).
- **Chain-of-Thought (CoT):** Introduce reasoning methods to structure multi-turn dialogues, enhancing the model’s ability to reason through complex tasks.

5.6 Summary

Through experimental validation, the proposed method demonstrated substantial performance improvements in multimodal intent recognition:

- After introducing OCR training and inference, the model achieved SOTA with an average score of 83.98.
- With LoRA parameter optimization (rank=16, scaling factor=32), the model achieved the best result of 84.17.

Future work will focus on further optimizing OCR text processing, refining parameters, integrating RAG and Chain-of-Thought techniques, and enhancing the model’s generalization and robustness in more complex scenarios.

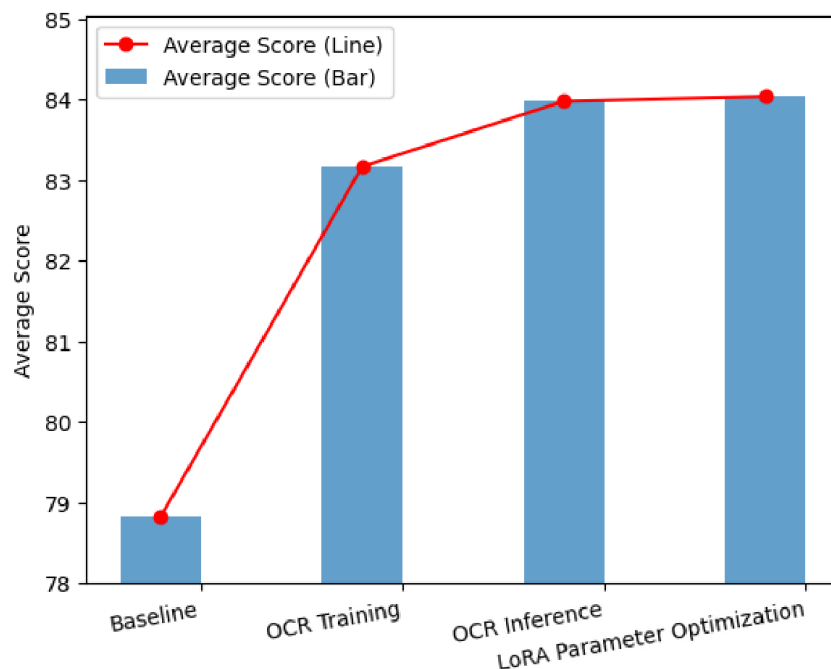


Figure 7: Performance improvement with different configurations, showing the increase from baseline (78.82) to the best result (84.17).

6 Conclusion

This study addresses the challenge of intent recognition in multimodal dialogue systems within the e-commerce domain by proposing an innovative method based on large language models (LLMs).

As user expressions on e-commerce platforms become increasingly diverse, incorporating multiple input modalities such as text, voice, and images, accurately understanding user intent and responding efficiently has become crucial for enhancing user experience and optimizing system performance. However, existing mainstream large language models, despite their excellent multitask learning capabilities in open domains, still exhibit significant shortcomings in domain-specific tasks, particularly in complex e-commerce dialogue scenarios. Issues such as inaccurate intent recognition, severe information omission, and poor context correlation not only degrade user experience but also impact customer satisfaction and sales conversion rates on e-commerce platforms.

To overcome these challenges, this study proposes fine-tuning advanced large language models (e.g., Qwen2-VL) using Low-Rank Adaptation (LoRA) to better adapt to the specific needs of the e-commerce domain. LoRA, as an efficient fine-tuning method, introduces low-rank matrices to adjust model parameters without requiring full retraining, significantly enhancing model performance on domain-specific tasks. Additionally, to address the limitations posed by insufficient data and modality heterogeneity, this study incorporates a variety of data augmentation techniques. These techniques include multiple enhancement strategies for both text and images. For text data, methods such as synonym replacement, syntactic reordering, and random deletion of key characters are used to help the model handle the diversity of user expressions. For image data, techniques like cropping, rotation, color adjustments, and noise addition are employed, along with Optical Character Recognition (OCR) to extract key information from images. These methods effectively mitigate information loss during image-text fusion, expanding the dataset from hundreds to thousands of samples and improving the model’s robustness and generalization ability.

Inspired by knowledge distillation and Retrieval-Augmented Generation (RAG) techniques, this study further integrates large language models with external knowledge bases. Knowledge distillation optimizes the model’s learning process through a teacher-student framework, allowing lightweight models to retain high performance while reducing computational overhead. RAG technology dynamically retrieves external information (e.g., product descriptions, user reviews, and frequently asked questions) during dialogue generation, providing contextual support for the model’s reasoning, which improves both the accuracy of intent recognition and the precision of responses. This multimodal fusion of information, combined with knowledge expansion, enables the model to more effectively capture complex user intents in e-commerce scenarios.

Through rigorous ablation experiments and parameter tuning, this study validates the effectiveness of the proposed method. Specifically, by combining OCR inference with parameter adjustments, the model’s average performance improved by 5.35 percentage points, compared to the baseline method. Both intent recognition and image scene understanding scores showed stable gains. These results demonstrate that fine-tuning large language models effectively, while integrating multimodal data and external knowledge, significantly enhances the intent recognition capabilities of multimodal dialogue systems in the e-commerce context.

In conclusion, this study proposes a multimodal intent recognition method tailored to the e-commerce domain, addressing the shortcomings of general models through fine-tuning and data augmentation strategies. Future work will focus on optimizing OCR result filtering mechanisms, exploring the integration of RAG and Chain-of-Thought reasoning, and further improving the intelligence and robustness of multimodal dialogue systems to provide more precise, personalized services and commercial value for e-commerce platforms.

References

- [1] H. Chen et al. “A survey on dialogue systems: Recent advances and new frontiers”. In: *ACM SIGKDD Explorations Newsletter* 19.2 (2017), pp. 25–35.
- [2] E. D. Cubuk et al. “Randaugment: Practical automated data augmentation with a reduced search space”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2020, pp. 702–703.
- [3] G. Hinton. “Distilling the Knowledge in a Neural Network”. In: *arXiv preprint arXiv:1503.02531* (2015).
- [4] A. Koschan and M. Abidi. *Digital color image processing*. John Wiley & Sons, 2008.
- [5] P. Lewis et al. “Retrieval-augmented generation for knowledge-intensive NLP tasks”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9459–9474.

- [6] J. Ni et al. “Recent advances in deep learning based dialogue systems: A systematic survey”. In: *Artificial Intelligence Review* 56.4 (2023), pp. 3055–3155.
- [7] D. Ramachandram and G. W. Taylor. “Deep multimodal learning: A survey on recent advances and trends”. In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 96–108.
- [8] C. Shorten, T. M. Khoshgoftaar, and B. Furht. “Text data augmentation for deep learning”. In: *Journal of Big Data* 8.1 (2021), p. 101.
- [9] A. Singh, K. Bacchuwar, and A. Bhasin. “A survey of OCR applications”. In: *International Journal of Machine Learning and Computing* 2.3 (2012), p. 314.
- [10] A. X. Yang et al. “Bayesian Low-Rank Adaptation for Large Language Models”. In: *arXiv preprint arXiv:2308.13111* (2023). URL: <https://arxiv.org/abs/2308.13111>.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/383433473>

Exploring Causal Inference in E-commerce Sentiment Analysis: Benchmarking LLMs for Customer Feedback Insights

Research · August 2024
DOI: 10.13140/RG.2.2.30537.56160

CITATIONS
0

READS
71

2 authors, including:



Marcus Trescothick
84 PUBLICATIONS 1,790 CITATIONS

SEE PROFILE

Exploring Causal Inference in E-commerce Sentiment Analysis: Benchmarking LLMs for Customer Feedback Insights

Authors: Nick Knight, Marcus Trescothick

Abstract

In the evolving landscape of e-commerce, leveraging customer feedback effectively is crucial for driving business success and enhancing customer satisfaction. Traditional sentiment analysis techniques, while valuable, often provide a superficial understanding of customer opinions without addressing the underlying factors influencing these sentiments. This study explores the integration of causal inference with Large Language Models (LLMs) to offer deeper insights into customer feedback. By benchmarking LLMs against conventional sentiment analysis methods, this research aims to evaluate their effectiveness in identifying causal relationships within customer feedback. LLMs, such as GPT-4, possess advanced natural language processing capabilities that enable them to interpret complex language patterns and nuances with high accuracy. This study assesses how LLMs enhance sentiment analysis by providing a more contextual and nuanced understanding of customer reviews compared to traditional methods. Moreover, the research investigates how causal inference techniques, when combined with LLMs, can uncover the specific factors driving customer satisfaction or dissatisfaction. For instance, it identifies whether negative feedback is primarily due to product defects, delivery issues, or customer service failures. The integration of multimodal data—combining text with numerical ratings and behavioral metrics—is also explored to provide a holistic view of customer sentiment. This approach allows for a more comprehensive analysis, revealing patterns and correlations that might not be evident from text alone. The findings indicate that LLMs, supported by causal inference and multimodal analysis, offer significant improvements over traditional sentiment analysis methods. They provide more accurate, actionable insights into customer feedback, enabling businesses to address root causes of dissatisfaction and implement more effective strategies. This approach allows businesses to gain deeper insights, make data-driven decisions, and enhance overall customer satisfaction.

Keywords: Causal Inference, E-commerce, Sentiment Analysis, Large Language Models, LLMs, Customer Feedback, Benchmarking, Multimodal Data, GPT-4, Data-Driven Insights.

Introduction

In the dynamic realm of e-commerce, understanding and leveraging customer feedback is essential for driving business improvements and maintaining competitive advantage. Traditional sentiment analysis has been instrumental in gauging customer opinions, but it often falls short in providing a comprehensive understanding of the underlying factors influencing these sentiments. While sentiment analysis can classify feedback into categories such as positive, negative, or neutral, it generally lacks the depth needed to uncover the root causes of customer satisfaction or dissatisfaction. Large Language Models (LLMs), such as GPT-4, offer advanced capabilities that enhance sentiment analysis by interpreting complex language patterns and contextual nuances with greater accuracy. Unlike traditional methods that rely on predefined lexicons and simple algorithms, LLMs are trained on vast and diverse datasets, enabling them to understand subtle variations in language and sentiment. This advanced capability allows LLMs to provide a more detailed and nuanced view of customer feedback, capturing sentiments that may be missed by conventional models. However, to truly understand the factors driving customer sentiments, causal inference is required [1], [2]. Causal inference involves identifying and analyzing the cause-and-effect relationships within data. In the context of e-commerce, this means determining which factors—such as product quality, delivery times, or customer service—most significantly influence customer opinions. By integrating causal inference with sentiment analysis, businesses can gain insights into the specific reasons behind customer feedback, allowing them to address issues more effectively.

The integration of LLMs with causal inference also benefits from multimodal analysis, which involves combining textual feedback with other data types, such as numerical ratings and behavioral metrics. This approach offers a more comprehensive view of customer experiences by correlating different forms of data. For example, combining text reviews with star ratings can reveal whether negative sentiments are associated with particular product features or service issues. This study explores the potential of integrating LLMs with causal inference techniques to enhance the understanding of customer feedback in e-commerce. By benchmarking LLMs against traditional sentiment analysis methods, the research evaluates their effectiveness in providing deeper insights and uncovering causal relationships within customer feedback. The aim is to demonstrate how advanced models and techniques can lead to more accurate, actionable insights

that drive business improvements and enhance customer satisfaction. This integrated approach allows businesses to move beyond basic sentiment classification to gain a thorough understanding of the factors influencing customer opinions, ultimately leading to more informed decision-making and improved customer experiences [3].

Advancements in Sentiment Analysis with LLMs

Sentiment analysis has long been a cornerstone of understanding customer feedback, helping businesses gauge the emotional tone of reviews and comments. Traditional sentiment analysis methods, which often rely on rule-based systems or simple statistical approaches, have provided valuable insights but are limited in their ability to capture the full complexity of customer sentiments. Large Language Models (LLMs), such as GPT-4, represent a significant advancement in this field, offering a more sophisticated and nuanced approach to analyzing text data.

Limitations of Traditional Methods

Traditional sentiment analysis methods typically use predefined sentiment lexicons and rule-based algorithms to classify text into categories like positive, negative, or neutral. While these approaches can be effective for straightforward feedback, they often struggle with the subtleties of human language. Sarcasm, irony, and context-specific meanings can easily lead to misinterpretation. For example, a comment like "This is just perfect" might be interpreted as positive by traditional systems but could be sarcastic depending on the context. Such limitations can result in inaccurate sentiment classifications and a lack of understanding of the true customer experience.

Enhanced Contextual Understanding with LLMs

Large Language Models, trained on vast datasets of diverse text, offer a more advanced approach to sentiment analysis. Unlike traditional methods, LLMs are capable of understanding the context in which words and phrases are used, allowing them to capture nuances and subtleties that simpler models might miss [4]. For instance, LLMs can interpret complex sentences and understand the sentiment behind statements that involve mixed emotions or ambiguous language. This capability results in more accurate sentiment classifications and a deeper understanding of customer feedback.

Advanced Capabilities of GPT-4

GPT-4 and similar LLMs utilize deep learning techniques to analyze and generate text with high contextual awareness. These models are trained on extensive corpora, enabling them to understand a wide range of language patterns and expressions. This advanced capability allows LLMs to distinguish between genuine praise and sarcastic remarks, providing a more precise assessment of customer sentiment. Additionally, LLMs can handle lengthy and complex feedback more effectively, ensuring that nuanced sentiments are captured accurately.

Applications in E-commerce

In the e-commerce sector, the enhanced capabilities of LLMs in sentiment analysis can significantly impact business decision-making. For example, LLMs can provide detailed insights into customer reviews by identifying specific aspects of a product or service that are frequently mentioned in positive or negative contexts [5]. This information can help businesses understand which features are valued by customers and which aspects may need improvement. Furthermore, LLMs can analyze large volumes of feedback more efficiently than traditional methods, allowing companies to keep up with the rapid pace of online reviews and comments.

Benchmarking LLMs Against Traditional Methods

To assess the effectiveness of LLMs in sentiment analysis, it is essential to benchmark their performance against traditional methods. Metrics such as accuracy, precision, recall, and F1 score can be used to evaluate how well LLMs classify sentiment compared to rule-based systems and statistical models. Studies have shown that LLMs consistently outperform traditional methods, particularly in handling complex and nuanced language, providing businesses with more reliable and actionable insights.

Causal Inference: Uncovering Root Causes of Sentiments

While sentiment analysis provides valuable insights into the emotional tone of customer feedback, it often lacks the depth required to understand the underlying causes of these sentiments. Causal inference, a technique used to identify and analyze cause-and-effect relationships, addresses this gap by uncovering the specific factors that drive customer opinions. By integrating causal

inference with sentiment analysis, businesses can gain a more comprehensive understanding of customer feedback, leading to more targeted and effective strategies for improvement.

The Role of Causal Inference

Causal inference involves identifying and understanding the relationships between variables to determine which factors influence certain outcomes. In the context of e-commerce, this means investigating how various elements—such as product features, pricing, delivery times, or customer service—affect customer sentiments [6]. For instance, if a significant number of customers leave negative reviews, causal inference can help pinpoint whether this dissatisfaction is due to product defects, delays in shipping, or poor customer service interactions.

Challenges in Traditional Analysis

Traditional methods of causal analysis often rely on structured data and controlled experiments to establish cause-and-effect relationships. However, applying these methods to unstructured text data, such as customer reviews, presents challenges. The complexity of natural language, combined with the interplay of multiple factors influencing sentiments, makes it difficult to isolate specific causes. For example, a negative review might mention several issues simultaneously, making it challenging to determine which factor is most influential. Additionally, traditional statistical techniques may struggle to capture the nuances of customer feedback, leading to incomplete or inaccurate conclusions.

Leveraging LLMs for Causal Inference

Large Language Models (LLMs) offer a solution to these challenges by enabling advanced processing and analysis of unstructured text data. By combining LLMs with causal inference techniques, businesses can gain deeper insights into the factors driving customer sentiments. For example, LLMs can analyze customer reviews to identify patterns and correlations between specific complaints and overall sentiment trends. This capability allows businesses to determine whether recurring issues—such as "long delivery times" or "product quality problems"—are the primary drivers of negative feedback.

Applications in E-commerce

The application of causal inference, enhanced by LLMs, provides numerous benefits for e-commerce businesses. By uncovering the root causes of customer dissatisfaction, companies can implement targeted improvements in areas such as product design, supply chain management, and customer service [7]. For instance, if causal analysis reveals that delayed deliveries are a major cause of negative sentiment, a business might prioritize optimizing its logistics processes. Similarly, if product quality issues are identified as a significant factor, efforts can be focused on enhancing quality control measures.

Combining Causal Inference with Sentiment Analysis

Integrating causal inference with sentiment analysis offers a more holistic understanding of customer feedback. Sentiment analysis reveals what customers feel, while causal inference explains why they feel that way. This combination enables businesses to not only assess the overall sentiment but also understand the specific factors contributing to it. By addressing the root causes identified through causal inference, companies can develop more effective strategies to enhance customer satisfaction and loyalty.

Multimodal Analysis: Integrating Text, Ratings, and Behavioral Data

In the quest to fully understand customer feedback in e-commerce, relying solely on text-based sentiment analysis or numerical ratings can provide an incomplete picture. To gain a more comprehensive view, multimodal analysis—combining various data types such as textual reviews, star ratings, and behavioral metrics—offers a richer and more nuanced understanding of customer experiences [8]. By leveraging Large Language Models (LLMs) in conjunction with multimodal data, businesses can uncover deeper insights and make more informed decisions.

Understanding Multimodal Analysis

Multimodal analysis involves the integration and analysis of multiple types of data sources. In e-commerce, this can include text reviews, numerical ratings (e.g., star ratings), and behavioral data (e.g., click-through rates, purchase history). Each data type provides different insights into customer experiences. For example, while textual reviews can offer detailed qualitative feedback, star ratings provide a quantitative measure of overall satisfaction, and behavioral data reveals

patterns in customer interactions. Combining these data types allows for a more holistic view of customer sentiment and behavior.

Challenges with Single-Modal Analysis

Single-modal analysis—whether based solely on text or numerical ratings—has inherent limitations. Text-based sentiment analysis may miss quantitative trends or fail to capture the overall sentiment conveyed by star ratings. Conversely, numerical ratings alone can lack the contextual details needed to understand the reasons behind high or low scores. For instance, a product might receive a low rating due to issues not explicitly mentioned in text reviews, such as subtle defects or delays in service. Similarly, positive reviews might not fully capture underlying concerns or minor complaints that affect overall satisfaction.

LLMs in Multimodal Analysis

Large Language Models (LLMs) enhance multimodal analysis by effectively processing and integrating textual data with other data types. LLMs, such as GPT-4, are capable of understanding and correlating information from diverse sources. For example, an LLM can analyze text reviews to identify recurring issues and then cross-reference these findings with numerical ratings to determine whether specific problems correlate with lower scores [9]. Additionally, LLMs can integrate behavioral data to understand how sentiment trends impact customer actions, such as repeat purchases or product returns.

Applications in E-commerce

The application of multimodal analysis in e-commerce can provide valuable insights for improving customer satisfaction and operational efficiency. For instance, by combining text reviews with star ratings, businesses can identify specific product features or service aspects that consistently affect customer satisfaction. Multimodal analysis can also reveal patterns in behavioral data, such as whether negative feedback is associated with high rates of product returns or cart abandonment. This comprehensive view enables businesses to address root causes of dissatisfaction and make targeted improvements.

Benchmarking Multimodal Approaches

Benchmarking multimodal approaches against traditional single-modal methods helps assess their effectiveness. Metrics such as correlation accuracy, insight depth, and overall relevance of findings can be used to evaluate how well multimodal analysis captures and integrates diverse data sources. Studies often show that multimodal approaches provide a more accurate and detailed understanding of customer sentiment, leading to better-informed business decisions. Multimodal analysis, enhanced by LLMs, represents a significant advancement in understanding customer feedback in e-commerce. By integrating textual reviews, numerical ratings, and behavioral data, businesses can gain a comprehensive and nuanced view of customer experiences. This holistic approach allows for more accurate insights, targeted improvements, and ultimately, enhanced customer satisfaction. As e-commerce continues to grow, the ability to effectively analyze and leverage multimodal data will be crucial for maintaining a competitive edge and driving long-term success [10].

Conclusion

The integration of Large Language Models (LLMs) with causal inference and multimodal analysis marks a transformative advancement in e-commerce sentiment analysis. Traditional methods, while useful, often fall short in capturing the depth and complexity of customer feedback. LLMs offer enhanced capabilities in understanding and interpreting nuanced language, addressing the limitations of simpler sentiment analysis techniques. By employing advanced models like GPT-4, businesses can achieve a more accurate and contextually aware analysis of customer sentiments. Causal inference adds another layer of depth by uncovering the root causes behind customer feedback. This technique allows businesses to identify specific factors—such as product defects, delivery issues, or service failures—that influence customer opinions. By combining causal inference with LLMs, companies can gain a clearer understanding of why customers feel a certain way, leading to more targeted and effective interventions. Multimodal analysis further enriches this understanding by integrating diverse data sources, such as textual reviews, numerical ratings, and behavioral metrics. This comprehensive approach provides a more complete picture of customer experiences, revealing patterns and correlations that single-modal analysis might miss. For instance, analyzing text reviews in conjunction with star ratings and purchase behavior can help businesses identify which aspects of their products or services need improvement and how these issues impact overall satisfaction. The adoption of these advanced techniques enables e-

commerce businesses to move beyond basic sentiment classification and gain actionable insights that drive strategic decisions. By leveraging LLMs, causal inference, and multimodal analysis, companies can address the underlying causes of customer dissatisfaction, enhance their offerings, and ultimately improve customer loyalty and satisfaction. The integration of LLMs with causal inference and multimodal analysis represents a significant step forward in the field of e-commerce sentiment analysis. This approach not only provides a more nuanced and accurate understanding of customer feedback but also empowers businesses to make data-driven decisions that foster long-term success. As the e-commerce landscape continues to evolve, leveraging these advanced analytical techniques will be crucial for staying competitive and meeting the ever-changing needs of customers.

References

- [1] Jim, Jamin Rahman, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, and M. F. Mridha. "Natural Language Processing Journal."
- [2] Bao, Keqin, Jizhi Zhang, Xinyu Lin, Yang Zhang, Wenjie Wang, and Fuli Feng. "Large Language Models for Recommendation: Past, Present, and Future." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2993-2996. 2024.
- [3] Zeyu Wang, Yue Zhu, Shuyao He, Hao Yan, and Ziyi Zhu. 2024. "LLM for Sentiment Analysis in E-Commerce: A Deep Dive into Customer Feedback". *Applied Science and Engineering Journal for Advanced Research* 3 (4):8-13. <https://doi.org/10.5281/zenodo.12730477>.
- [4] Wang, Z. (2024, August). CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of Large Language Models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)* (pp. 143-151).
- [5] Bao, Keqin, Jizhi Zhang, Xinyu Lin, Yang Zhang, Wenjie Wang, and Fuli Feng. "Large Language Models for Recommendation: Past, Present, and Future." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2993-2996. 2024.

- [6] Wang, Cunxiang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao et al. "Survey on factuality in large language models: Knowledge, retrieval and domain-specificity." *arXiv preprint arXiv:2310.07521* (2023).
- [7] Syed, Ayesha Ayub, Ford Lumban Gaol, Alfred Boediman, and Widodo Budiharto. "Airline reviews processing: Abstractive summarization and rating-based sentiment classification using deep transfer learning." *International Journal of Information Management Data Insights* 4, no. 2 (2024): 100238.
- [8] Zhu, Yutao, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. "Large language models for information retrieval: A survey." *arXiv preprint arXiv:2308.07107* (2023).
- [9] Izadi, Saadat, and Mohamad Forouzanfar. "Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots." *AI* 5, no. 2 (2024): 803-841.
- [10] Chen, Zeyuan, Haiyan Wu, Kaixin Wu, Wei Chen, Mingjie Zhong, Jia Xu, Zhongyi Liu, and Wei Zhang. "Towards Boosting LLMs-driven Relevance Modeling with Progressive Retrieved Behavior-augmented Prompting." *arXiv preprint arXiv:2408.09439* (2024).



ChatSpot: Bootstrapping Multimodal LLMs via Precise Referring Instruction Tuning

Liang Zhao^{1*}, En Yu^{2*}, Zheng Ge^{1†}, Jinrong Yang², Haoran Wei¹,
Hongyu Zhou¹, Jianjian Sun¹, Yang Peng³, Runpei Dong⁴, Chunrui Han¹, Xiangyu Zhang¹

¹MEGVII Technology, ²Huazhong University of Science and Technology

³Tsinghua University, ⁴Xian Jiaotong University

Demo: <https://chatspot.streamlit.app>

Abstract

Human-AI interactivity is a critical aspect that reflects the usability of multimodal large language models (MLLMs). However, existing end-to-end MLLMs only allow users to interact with them through language instructions, leading to the limitation of the interactive accuracy and efficiency. In this study, we present *precise referring instructions* that utilize diverse reference representations such as points and boxes as referring prompts to refer to the special region. This enables MLLMs to focus on the region of interest and achieve finer-grained interaction. Based on precise referring instruction, we propose ChatSpot, a unified end-to-end multimodal large language model that supports diverse forms of interactivity including mouse clicks, drag-and-drop, and drawing boxes, which provides a more flexible and seamless interactive experience. We also construct a multi-grained vision-language instruction-following dataset based on existing datasets and GPT-4 generating. Furthermore, we design a series of evaluation tasks to assess the effectiveness of region recognition and interaction. Experimental results showcase ChatSpot’s promising performance.

1 Introduction

Recent advances in large language models (LLMs) exemplified by GPT-3 [3] and LLaMA [32] have demonstrated significant potential in the domain of zero-shot learning and logical reasoning. By aligning pre-trained LLMs to follow human language instructions through Reinforcement Learning with Human Feedback (RLHF) [7], InstructGPT [23] and ChatGPT [21] have showcased powerful capabilities for human-AI interaction, leading to a new paradigm shift towards the realization of artificial general intelligence (AGI).

Inspired by the remarkable success of GPT series [3, 21, 22], researchers attempt to incorporate more modalities into LLMs for multimodal human-AI interaction, with vision-language interaction being an important topic of focus. In order to incorporate visual modality into LLM, significant processes have been made to bridge the gap between LLMs and vision foundation models. There are two mainstream paradigms for building multimodal large language models (MLLMs). One is *plugin-based* MLLM [29, 35, 37] that utilizes off-the-shell LLMs [5, 21, 32] as central controllers to schedule different visual expert models as plugins. In this way, the users can interact with LLMs to

*Equal contribution

†Project leader

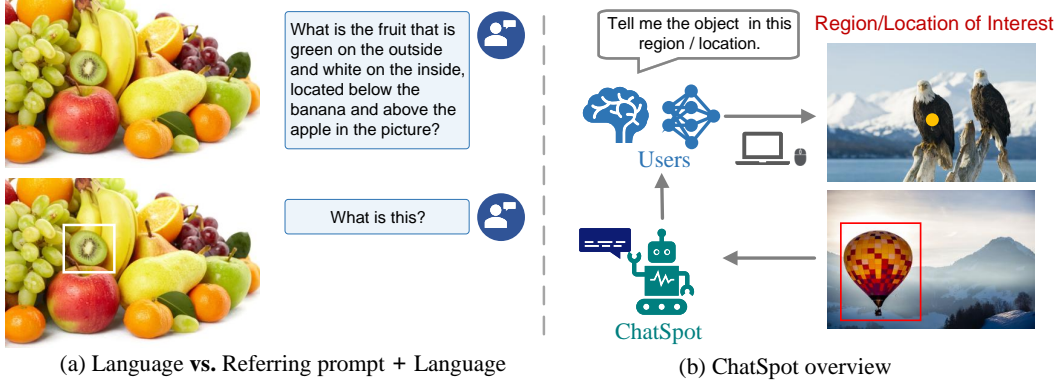


Figure 1: (a) is the intuitive comparison between language instruction and the combination of region prompt and language instruction. (b) is the overview of ChatSpot. We extend the power of advanced LLM to vision-language modality and support a range of interaction forms including native language, mouse-clicking, and mouse boxing, enabling the interaction to be more flexible and user-friendly.

achieve diverse visual functions. Another paradigm is *end-to-end* MLLM [1, 10, 15] that employs various techniques to align visual signals obtained from the vision encoder to the language semantic space and input vision tokens and language tokens together into the large language decoder.

Despite existing *end-to-end* MLLMs have achieved remarkable progress in vision-language human-AI interaction, the mode of interactive instruction is still limited in language. When meeting complex scenes as illustrated in Figure 1 (a), it is difficult to only use the language to accurately describe the requirement of the user. However, if we can add some **referring prompts**, *e.g.*, reference points, bounding boxes, *etc.*, to MLLMs, the model can focus on the region of interest (RoI) and achieve finer-grained interaction, which is more flexible and user-friendly.

Motivated by this, we present ChatSpot, a fully end-to-end unified multimodal language model designed to empower special region vision-language interaction. As illustrated in Figure 1 (b), ChatSpot extends the LLMs’ power to incorporate diverse multimodal inputs, and it can support a range of interaction forms. Users can communicate with the system using their native language, as well as gestures such as clicking and drawing boxes that we call **Precise Referring**, to obtain the desired information about the entire image or the region of interest (RoI). When a specific region is selected, ChatSpot can follow the precise referring instructions to perform various fine-grained applications, such as identifying jersey numbers or analyzing facial expressions in the given task, which is illustrated in Figure 4. Furthermore, the precise referring can be regarded as a link of the chain-of-thought (CoT) to enhance the special logical reasoning ability of MLLMs. When an intelligent agent (robot or expert model) locates a target or region of interest based on user demands, ChatSpot can further analyze the details of this region and provide more specific suggestions and refined instructions, enabling the agent to interact with the physical world more effectively.

The success of ChatSpot hinges on three components: (1) We design a simple but effective precise referring instruction tuning method for MLLMs to support fine-grained interaction. (2) We construct a high-quality Multi-grained vision-language instruction-following dataset (**MGVLID**) including image-text and region-text with around 1.2M images and 3M query-answer pairs by collecting from existing datasets and generating based on GPT-4. (3) We design a series of evaluation tasks and metrics to assess the effectiveness of the proposed model. Extensive experiments have been conducted on a wide of vision-centric and vision-language benchmarks, and our ChatSpot shows excellent performance.

2 Related Works

2.1 Large Language Models

In recent years, large language models (LLMs) have garnered considerable attention in the domains of natural language processing (NLP) and artificial general intelligence (AGI) owing to their remarkable performance in language generation, in-context learning, world knowledge, and logical reasoning.

Early works, *e.g.*, BERT [8], GPT-2 [25] and T5 [26] established the foundation architecture of LLMs. Then, with the release of GPT-3 [3], the first-ever language model to reach the parameter size of 175 billion, LLM achieved impressive zero-shot performance on various language benchmarks. Furthermore, researchers discovered *emergent ability* [34] in LLMs. That is when the model size of language models scales up to a certain level, there is a qualitative leap in the capabilities of language models. Sequentially, by aligning pre-trained GPT-3 to follow human language instructions through Reinforcement Learning with Human Feedback (RLHF) [7], InstructGPT [23] and ChatGPT [21] showcased powerful capabilities for human-AI interaction, which make LLMs reach its “iPhone moment”. Inspired by the great success of GPT series, many other open-sourced LLMs, such as OPT [40], LLaMA [32], and GLM [38], have been proposed, which achieve similar performance to GPT-3. Based on these open-sourced LLMs, several specific fine-tuned models are proposed to construct LLMs for various applications. For instance, Alpaca [31] proposes a self-instruct framework based on LLaMA [32] and employs 52K instructions generated by ChatGPT [21] to construct an exceptional dialogue model.

2.2 LLM-based MultiModal Interactive Agent

The success of LLMs [21, 32] may have opened the gate towards artificial general intelligence (AGI), a crucial component of which is human-AI interaction. The powerful zero-shot and logical reasoning ability of LLM makes it the central controller of the interactive system to schedule various application tools for different modality tasks, such as VQA, image editing, and image captioning. There are two mainstream interactive styles, *i.e.*, plugin-based and end-to-end interaction. Plugin-based methods [13, 29, 35–37] usually prompt LLM (ChatGPT [21], GPT-4 [22] or LLaMA [32]) to invoke different plugins from other foundation or expert models to perform specific functions according to human instructions. However, despite plugin-based methods that enable diverse applications, they are limited in the effectiveness of plugin invocation and the performance of the plugin model. On the contrary, end-to-end interactive systems usually use a single large multimodal model to accomplish interaction. This approach takes advantage of cross-modal transfer, aligning multimodal domains to a common language semantic space and then using autoregressive language models as decoders to output the language. Following this pipeline, Flamingo [1] developed a gated cross-attention trained on billions of image-text pairs to align vision and language modality, which shows strong performance in few-shot learning. BLIP-2 [12] introduced Q-Former to align visual features with language space more effectively. More recently, LLaVA [15] proposed to use a simple linear layer to replace Q-Former and design a two-stage instruction-tuning procedure. Although existing end-to-end methods achieve remarkable performance in high efficiency, they are all limited to the interaction form of the full image and language-only instruction, which can not satisfy the demand for the specific region interaction. In this work, we build an end-to-end unified multimodal language model that supports a range of interaction forms that supports both full images and specific region.

3 Methods

ChatSpot is a multimodal large language model capable of perceiving real-world multimodal information, as well as following instructions, reasoning, and interacting with humans in natural language. It supports diverse forms of interaction including natural language, mouse-clicking, and mouse boxing. In this work, we mainly consider the modalities of image and language. And we will support more diverse interaction modalities and forms, *e.g.*, video and audio, in the future.

3.1 Overall Architecture

As illustrated in Figure 2, ChatSpot consists of an image encoder, and a decoder-only LLM, and a modality alignment block. Inspired by LLaVA [15], ChatSpot incorporates a simple multilayer perceptron (MLP) to align the visual tokens with the space of language. The overall architecture is simple and does not use any extra AI models or post-processing operations. Different from previous end-to-end interactive systems [15, 41] that only support full image interaction, ChatSpot presents a more flexible interaction that supports users in further selecting the region of interest (RoI) to issue finer-grained instructions.

When an image I is uploaded by users to the ChatSpot system, users can use a mouse to select the RoI (points or boxes) R_t through a series of gestures, such as clicking and drawing boxes, and give some

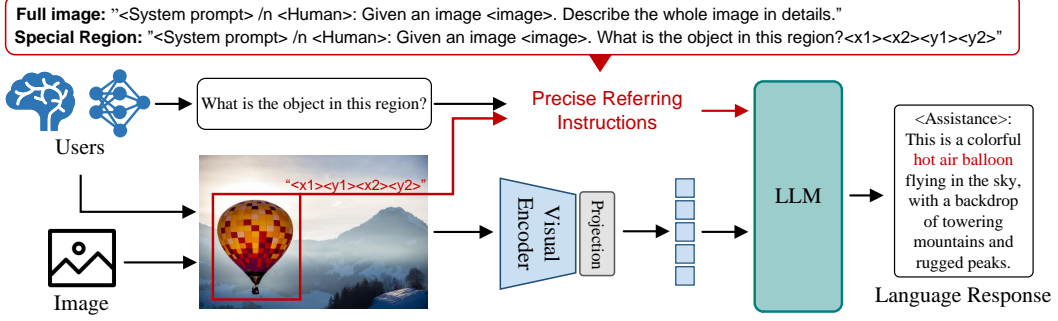


Figure 2: **Overall pipeline of ChatSpot.** The architecture of ChatSpot consists of three main components: (1) an image encoder, (2) a large language model, and (3) a modality-align projector.

language instructions X^{instruct} about these RoIs. The system then converts the position of R_t into the region prompt and connects it with the X^{instruct} to generate the precise referring instructions. Afterward, the image I is inputted into the visual encoder to extract visual tokens. And then the modality-align projector transforms the visual tokens to the language semantic space. After obtaining the refined visual tokens and precise referring instructions, LLM decoder \mathcal{F} takes them as inputs and generates the response language sequence Y autoregressively. Formally,

$$\begin{aligned} V &= \zeta \circ \mathcal{G}(I), \\ Y &= \mathcal{F}(V, \Phi(R_t), X^{\text{instruct}}), \end{aligned} \quad (1)$$

where V are aligned visual tokens. \mathcal{G} is the visual encoder and ζ denotes the vision-language alignment projection. \mathcal{F} is the large language decoder. $\Phi(\cdot)$ is the normalization operation.

3.2 Precise Referring Instruction

Due to the inherent semantic unit mismatch between images and texts, it is ineffective to directly use the whole image and language sentences to describe the vision-language task. To this end, we propose *precise referring instruction* that enables the unification of multi-grained vision-language task descriptions and supports proxy interaction forms. Specifically, we divide the instructions into two types, *i.e.*, image-level instructions and region-level instructions.

Image-level Instructions. The image-level instructions are usually used to describe the task of the whole image, and existing multimodal instructions mostly adopt this form. For instance, given an image and we want to know what the content of the image is. Then the instruction can be like “Given an image $\langle \text{image} \rangle$. Describe the whole image in detail”, or “Given an image $\langle \text{image} \rangle$, please tell me: $\langle \text{question} \rangle$ ”, where $\langle \text{image} \rangle$ is the input image and $\langle \text{question} \rangle$ denote some relative questions about images. Afterward, the LLM ingests the whole sentence and outputs the corresponding response.

Region-level Instructions. Compared to the overall information about the whole image, we often pay more attention to the information in specific regions. Therefore, it is valuable to design effective region-level instructions for multimodal large language models. The key challenge of region-level instructions is how to make LLM aware of the specific location of the region of interest. Here, we provide a simple but effective instruction format to achieve it. Specifically, we first define a unified region representation format as a tuple $R_t = \{x_k, y_k\}_{k=1}^N$ that represents N points located in the selected region. Then the coordinates of selected points are normalized to $[0, 1]$ and transferred to the text tokens as $\Phi(R_t)$. Finally, the region coordinate tokens are connected with the language instructions to generate the final region-level instructions. ([15] and [4] have served as evidence that LLM possesses the capability to comprehend spatial relationships and coordinates based on textual descriptions.) A simple example of region-level instruction is as follows: “Given an image $\langle \text{image} \rangle$, What is the object doing in the region? $\langle \text{region} \rangle$ ” where $\langle \text{region} \rangle = \langle \text{box} \rangle \Phi(R_t) \langle \text{box} \rangle$, $\langle \text{box} \rangle$ and $\langle \text{box} \rangle$ are special tokens to tell the LLM that this is a set of coordinates of the RoI.

Notably, the number of selected points N is set freely so that we can achieve multi-grained interaction, such as points, boxes, and polygons.

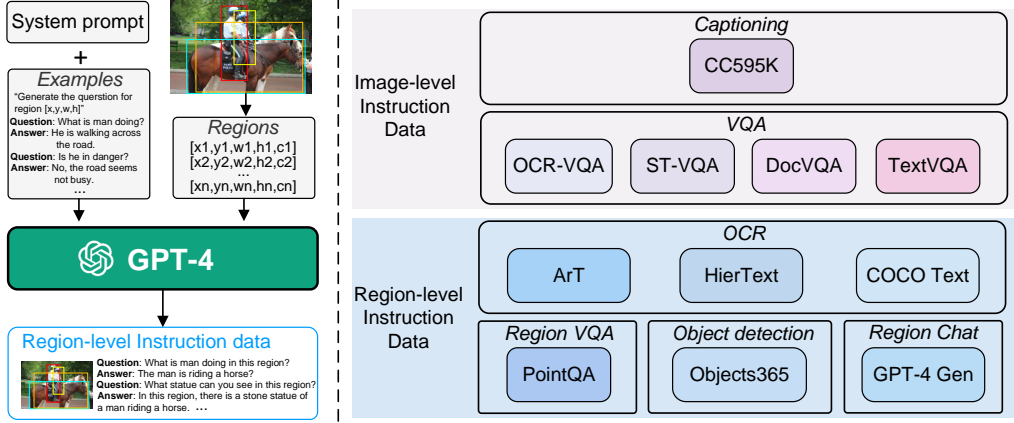


Figure 3: Illustration of the pipeline to collect region-level chatting data for MGVLID (left) and dataset groups included in Multigrained Vision-language Instruction Datasets, **MGVLID** (right).

3.3 Multi-grained Vision-language Instruction-following Dataset

In order to empower ChatSpot with the precise referring instruction following ability, we need to construct enough high-quality region-level instruction datasets. However, existing vision-language datasets lack diverse region-level chatting data. To this end, we design a data collection pipeline with the assistance of GPT-4 [22], as shown in Figure 3. Inspired by LLaVA [15], we use the captions and bounding boxes of the target as the prompts and leverage GPT-4 to refine these captions and generate more relative and diverse conversation data. The difference is that our approach distinguishes itself by enforcing the alignment of every generated dialogue with precise regional coordinates. To achieve this, we leverage the VisualGenome dataset [11], which provides comprehensive annotations of objects, attributes, and relationships within each image, enabling us to construct region-level instruction following datasets. These datasets consist of dense region-wise captions organized alongside carefully curated seed examples, which are used to query GPT-4 in an in-context-learning fashion. Through this pipeline, we have successfully gathered a total of 108K region-level instruction following samples.

Based on this data generation pipeline, we build a high-quality **Multi-Grained Vision-Language Instruction-following Dataset**, named **MGVLID**. MGVLID consists of two main parts, *i.e.*, image-text instruction-following data and region-text instruction-following data. The former data consists of the image and the caption of the entire image, while the latter consists of the image, the bounding boxes of the target in the image, and the corresponding target captions. As shown in Figure 3, the whole MGVLID covers 11 source datasets and we hold out 4 datasets for model evaluation purposes.

Image-level Instruction Data. To gather image-level instruction-following data, we collect a wide range of publicly available multimodal datasets that have been human-annotated. We then transform these datasets into a unified instruction-following format. Specifically, we assemble a plethora of commonly used Question-Answering (QA), captioning, and object detection datasets, including CC595K (filtered based on CC3M [28]), OCR-VQA [20], ST-VQA [2], DocVQA [19], TextVQA [30] and Object365 [27]. For each dataset, we design a series of unique instruction templates. These templates are subsequently carefully filtered and refined manually to ensure optimal rationales and diversity of the conversation. Due to the considerable differences in label lengths among various task datasets (such as caption or category words), we incorporate additional instruction tags to specify the desired response style. For instance, we include tags like “*answer in shot*” for short-answer data and “*answer in detail*” for long-answer data.

Region-level Instruction Data. While image-level instruction data empowers the model’s global visual perception and human instruction-following ability, region-level instruction data offers region-level observation and more fine-grained instructions, enabling the model to further acquire spatial perception and reasoning abilities. In order to construct region-level instruction datasets, we first collect region-text pairs based on existing region-level task (object detection and OCR) datasets, *i.e.*, Object365 [27], COCO text [33], HierText [16] and Art [6]. We collected region-text pairs that

consist of instance-level bounding boxes and their corresponding content. Subsequently, we utilize unique instruction templates to further refine these region-text pairs, resulting in a series of questions and answers. Furthermore, we also collect the PointQA datasets from LookTwice-QA [18] to support point-wise referring instruction tuning, where the models are asked to answer questions based on the input points or boxes. By incorporating the high-quality dense region chatting data generated based on GPT-4, the final region-level instruction data is constructed.

4 Experiments

4.1 Implementation Details

To build ChatSpot, we implement the CLIP ViT-L/14 [24] as the visual encoder to encode images. For the large language model, we choose open-sourced Vicuna-7B [5] as the language decoder, a LLaMA model fine-tuned with instructions. For alignment projection, we just adopt a simple linear layer to connect vision and language embedding space.

Inspired by LLaVA, the model is trained in a two-stage fashion. Firstly, we initialize the model using pre-trained weights from LLaMA and CLIP ViT. During this first stage, we only train the projection layer. Meanwhile, we freeze the majority of the LLM parameters. In this stage, we mainly use the image-text instruction-following data of MGVLID to train the model for vision-language instruction-following alignment learning. In the second stage, we only freeze the visual encoder and unfreeze the LLM parameters. In this stage, we mainly use the region-text instruction-following data including RegionChat to train the model for region-level instruction-following and multi-turn chatting ability. Specifically, the model is fine-tuned over 3 epochs, with a batch size of 128. AdamW [17] optimizer is employed, and the learning rate is set to $2e - 3$ in the first training stage and $2e - 5$ in the second training stage. For LLM, the maximum length of tokens is set to 2,048.

4.2 Task Evaluation

In order to objectively showcase ChatSpot’s region recognition and zero-shot ability. We choose several downstream tasks including regional classification, OCR text recognition, and VQA answer grounding tasks to evaluate ChatSpot. The results are shown in Table 1 and Table 2. Notably, all the experiments of different tasks are conducted by the shared-parameter generalist model, and we just change the language instructions for different tasks.

Regional Classification. Object detection is a fundamental vision task that consists of object location and recognition subtasks. In this part, we mainly evaluate the regional classification ability of ChatSpot in COCO [14], which is a common dataset in object detection tasks. Specifically, we first use the GT boxes or the bounding boxes generated by existing detectors, such as DINO [39], as region prompts to ask ChatSpot to answer what category it is. For an example, we use “*What can you see in this region? <region>*”, where *<region>* denotes the coordinates of the region boxes. Then we compute the metrics of Average Precision (AP) and Accuracy about the bounding boxes with the predicted classes. Notably, due to ChatSpot’s output typically being a single sentence, it cannot be directly used as a category for evaluation. Here, we employ the CLIP text encoder to calculate the text feature similarity between the output sentence and all COCO categories for category determination.

As shown in Table 1, we randomly select 1,000 images from the COCO validation set, namely COCO-1000, to evaluate ChatSpot for efficiency. Our ChatSpot achieves 64.5% accuracy on COCO-1000 with the provided GT boxes. When given DINO-generated bounding boxes as region prompts, our ChatSpot achieves 39.6% AP, which is also a competitive performance. Notably, We do not use any annotations from COCO. Therefore, the results show that ChatSpot achieves impressive zero-shot classification ability in region-level recognition.

Regional Optical Character Recognition. Optical Character Recognition (OCR) is a visual entity recognition task that requires the recognition of the graphemes in a written text. In this part, we select COCO text [33] to evaluate the regional text recognition ability of ChatSpot. COCO Text is a large-scale dataset for text detection and recognition. We first use the provided region boxes of the datasets as the region referring and ask the ChatSpot “*What text can you see in this region? <region>*”. Then ChatSpot will respond to the specific answer. Similar to evaluating VizWiz, when the answered sentence includes the correct GT answer, we consider ChatSpot’s response to be correct. As shown in Table 2, our ChatSpot achieves 31.8% accuracy on the COCO Text validation set.

Table 1: **Zero-shot region recognition results on COCO val set.** We randomly select 1,000 images from the COCO validation set for evaluation. The referring regions are provided by GT boxes and advanced detector DINO [39]. Acc. denotes the classification accuracy when given the GT boxes.

Method	Backbone	Training Data	Region	COCO [14]						
				AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l	Acc.
<i>Multi-modal Large Language Models</i>										
ChatSpot	CLIP-ViT	MGVLID	GT boxes	48.8	48.8	48.8	35.1	56.0	60.3	64.5
ChatSpot	CLIP-ViT	MGVLID	DINO boxes	39.6	50.2	44.1	21.6	45.8	58.8	-

Table 2: **Experimental results on a diverse set of downstream tasks.** We also evaluate ChatSpot on a series of downstream tasks including optical character recognition (OCR) and visual question answering (VQA). We mainly report the metric of Accuracy (%) for evaluation. For PointQA, “B” and “P” mean that answer the question based on the given box and point, respectively.

Method	Backbone	Training Data	OCR		VQA		
			COCO Text	VizWiz	PointQA (B)	PointQA (P)	
ChatSpot	CLIP-ViT	MGVLID	31.8	63.0	68.2	62.0	

Visual Question Answering (VQA). VQA is the task of answering open-ended questions based on the whole image or the region of interest (RoI), which is well-suited for evaluating the perceptual and reasoning abilities of large multimodal language models in understanding image content. In this part, we choose two datasets to evaluate ChatSpot. One is VizWiz-VQA-Grounding dataset [9], a dataset that visually grounds answers to visual questions asked by people with visual impairments. Another is PointQA [18], a set of datasets that require a pointer to an object in the image to be answered correctly. These two datasets both provide specific regions (boxes or points) and ask the question about the corresponding area. Therefore, it also requires the model to possess the ability of region-level perception and reasoning ability. Specifically, given a region of interest and the corresponding question, we first input the coordinate of region boxes to ChatSpot with the corresponding question. If the output answer sentence of ChatSpot includes the GT answer, we consider the response to be correct. As shown in Table 2, benefiting from ChatSpot’s strong region-level instruction following abilities, the model achieves competitive performance that obtains 63.0% accuracy on VizWiz and 68.2% accuracy on PointQA (boxes given) and 62.0% on PointQA (points given).

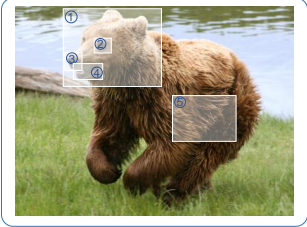
4.3 Qualitative Analysis

In order to provide a comprehensive showcase of ChatSpot, we selected several classic cases to demonstrate its specific abilities. We specifically demonstrate four core capabilities of ChatSpot through these examples as follows:

Region Perception Ability. As shown in Figure 4 (left top), we first show the case that depicts ChatSpot’s ability to perceive the region of interest and recognize the corresponding context. In this case, ChatSpot can identify the selected area of different levels of granularity, *e.g.*, the head of the brown bear and the nose of the brown bear. It can also perceive some specific features of the target like the wet hair of the bear. By collaborating with human referring prompts (point or boxes), ChatSpot showcases its powerful capability in perceiving details, which provides sufficient detailed information for robots to perform more refined operations.

Content Generation Capability. ChatSpot also possesses a powerful content generation capability related to regions of interest, as illustrated in Figure 4 (left-right). In this case, ChatSpot first recognizes the corresponding object (lemon) in the region of interest. Then it can also generate responses for other information that cannot be captured in the image, *i.e.*, the precautions when cutting a lemon with a knife. Thanks to the huge knowledge of LLMs, ChatSpot can provide rich content explanations or suggestions based on visual information.

Optical Character Recognition (OCR) Ability. ChatSpot also achieves impressive performance in recognizing the optical character, *e.g.*, text, number, and signal. As shown in Figure 4 (bottom left),



① What can you see in this area?

I can see the head of a brown bear.

② What is this?

This is the eye of a brown bear.

③ What about this?


This is the nose of a brown bear.

④ And this?

This is the mouse of a brown bear.

⑤ Is this bear's hair dry or wet?

This bear's hair appears to be wet.



① With this, what can you do in this image?

You can cut fruits with this knife.

② What fruits can you see in this region?


There are two pineapples in this region.

③ What about this?

This is a lemon in this region.

④ What are the things I should be cautious about when I cut this?

When cutting a lemon, be cautious of the acidic juice, which can cause a stinging sensation on the skin and may damage some materials. Also, be careful not to squeeze the lemon too hard, as it may cause the seeds to be released.



① What is the number shown here?

The number shown here is 20.

② What about this number?


This number is 18.

③ What about this number?

This number is 18.

④ What are these numbers used for in this scene?

These numbers are positioned above the checkout lanes in the store. They are used to indicate the order of the lanes, helping customers and store staff to identify the proper sequence for checkout.



① What is this?

This is a refrigerator.

② Is this fridge plugged in?

Yes, the fridge is plugged in. You can see the plug inserted into the outlet, and the wire connected to the fridge.

You are a robot. You want to take a Coke from this fridge. List all the actions you will take.

As a robot, I am not capable of taking anything from fridge. However, assuming the role of a human, I would need to open the fridge door, locate the Coke bottle, grab it by its handle, and then close the door to keep the cold air inside.

Figure 4: Cases of interacting with ChatSpot through drawing bounding boxes.

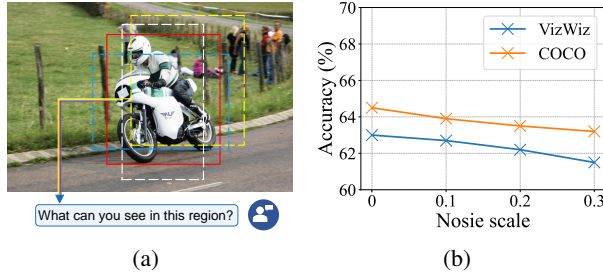


Figure 5: **Experiment on the robustness of region referring.** (a) demonstrates the process of randomly adding noise to the original region boxes and posing the question to ChatSpot. (b) showcases the performance of ChatSpot after incorporating random noise perturbations.

ChatSpot can accurately identify the number written on the signboard and analyze the purpose of these numbers based on the global context information.

Special Reasoning Ability. In addition to perception, another important capability of ChatSpot is spatial reasoning which can further analyze the region of interest based on its knowledge after recognition, which is important for robotics and automation. As shown in Figure 4 (bottom right), ChatSpot first identifies the fridge and then determines that the refrigerator is powered on according to the power wire being plugged into the power strip and connected to the refrigerator. Furthermore, ChatSpot provides a series of detailed action instructions when the user wants to take a Coke from the fridge. This enables the robots to be able to make further decisions regarding fine-grained operations after inferring the specific status (fridge is plugged in.) of the RoI.

5 Discussion

In this section, we are dedicated to delving deeper into the capabilities and features of ChatSpot, as well as identifying the limitations that currently hinder further enhancement of ChatSpot’s abilities.

5.1 Robustness on Region Referring

The process of selecting regions, whether through drawing boxes or clicking, plays a crucial role in ChatSpot. However, users often face difficulties in accurately annotating their areas of interest. In such instances, it is essential for ChatSpot to exhibit a high level of robustness in region selection. Hence, an analysis of the robustness of ChatSpot in the region referring is conducted. Specifically, we randomly introduce box noises of different scales (scale = 0.1, 0.2, and 0.3) to the region boxes of the COCO and VizWiz as illustrated in Figure 5 (a). Figure 5 (b) demonstrates that the performance of ChatSpot on COCO and VizWiz does not show a significant decrease after introducing box noises into the region bounding box, which means ChatSpot possesses strong robustness in region referring.

5.2 Region Referring Hallucination

There exists a potential risk that ChatSpot may mistakenly recognize the region referred to by users as a nearby region, and we coin this phenomenon as the “region referring hallucination”. To quantitatively assess whether ChatSpot is prone to region referring hallucination issues, we analyze the cases of misidentification in the sampled COCO-1000 dataset. Specifically, we define an occurrence of the “region referring hallucination” when an object of a certain class, misidentified by ChatSpot, appears in close proximity to the selected region, with an Intersection over Union (IoU) greater than 0.5.

As shown in Table 3, ChatSpot has shown very few instances of regional illusions, accounting for only about 2% of the total. Furthermore, even with the introduction of box noise, the illusion ratio does not have a significant change (2.2%-2.3%). This result demonstrates that ChatSpot exhibits a capability for precise region referring. The current recognition errors can be largely attributed to the model’s insufficient training on an adequate amount of data.

Table 3: **Results of region referring hallucination.** s is the noise scale. “Acc.” indicates the Accuracy and “Hallucination Ratio” denotes the proportion of region referring hallucination.

Model	Setting	COCO	
		Acc.	Hallucination Ratio
ChatSpot	No noise	64.7	2.2
	Box noise ($s = 0.1$)	63.9	2.4
	Box noise ($s = 0.2$)	63.5	2.1
	Box noise ($s = 0.3$)	63.2	2.3

5.3 Limitations

Although ChatSpot has achieved remarkable performance in precise region referring and special region reasoning, it still has some noticeable limitations. For example, ChatSpot currently lacks the ability to support referential output boxes and does not support the recognition of certain special symbols, such as license plate numbers. These shortcomings can be attributed to insufficient training data. Another limitation is the phenomenon of catastrophic forgetting, whereby fine-tuning ChatSpot on a new dataset leads to the forgetting of previously acquired knowledge, leading to the overall performance bottleneck. Furthermore, it is hard to evaluate the region recognition ability of multi-modal large language models since its evaluation is essentially different from traditional visual-language models. Though we take the first step to quantitatively evaluate it by giving pre-defined boxes, it is still an open problem: *how can we establish a comprehensive and automatic benchmark to evaluate existing multimodal large language models?* These limitations require further research from the community. ChatSpot currently only supports the interaction forms of mouse-clicking and drawing boxes. In the future, we will support more diverse interaction forms, *e.g.*, polygon, and mask.

6 Conclusion

In this work, we first propose *precise referring instruction* tuning for multimodal LLMs (MLLMs) that utilizes diverse reference representations for referring special regions. Based on precise referring instruction, we build ChatSpot, a fully end-to-end MLLM that supports diverse region referring prompts, *i.e.*, points, and boxes. Then we construct a large-scale multi-grained vision-language instruction-following dataset, MGVLID. Trained on MGVLID, ChatSpot demonstrates outstanding performance both in interactive chatting and downstream tasks. Results suggest that combining precise referring instructions with MLLMs stimulates the model’s ability for special region understanding and reasoning. We hope this work can spur more advanced MLLMs in the future.

References

- [1] Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022) 2, 3
- [2] Biten, A.F., Litman, R., Xie, Y., Appalaraju, S., Manmatha, R.: Latr: Layout-aware transformer for scene-text vqa. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16548–16558 (2022) 5
- [3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901 (2020) 1, 3
- [4] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023) 4
- [5] Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. <https://lmsys.org/blog/2023-03-30-vicuna/> (2023) 1, 6

- [6] Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., et al.: Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1571–1576. IEEE (2019) 5
- [7] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. *Advances in neural information processing systems* **30** (2017) 1, 3
- [8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018) 3
- [9] Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3608–3617 (2018) 7
- [10] Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al.: Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045* (2023) 2
- [11] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123**, 32–73 (2017) 5
- [12] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023) 3
- [13] Liang, Y., Wu, C., Song, T., Wu, W., Xia, Y., Liu, Y., Ou, Y., Lu, S., Ji, L., Mao, S., et al.: Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434* (2023) 3
- [14] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: *ECCV*. pp. 740–755 (2014) 6, 7
- [15] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 2, 3, 4, 5
- [16] Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1049–1059 (2022) 5
- [17] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *ICLR* (2019) 6
- [18] Mani, A., Yoo, N., Hinthorn, W., Russakovsky, O.: Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681* (2020) 6, 7
- [19] Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 2200–2209 (2021) 5
- [20] Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR). pp. 947–952. IEEE (2019) 5
- [21] OpenAI: Chatgpt. <https://openai.com/blog/chatgpt/> (2023) 1, 3
- [22] OpenAI: Gpt-4 technical report (2023) 1, 3, 5
- [23] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: *NeurIPS* (2022) 1, 3
- [24] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021) 6
- [25] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019) 3
- [26] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21**(1), 5485–5551 (2020) 3

- [27] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019) [5](#)
- [28] Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018) [5](#)
- [29] Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580 (2023) [1](#), [3](#)
- [30] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019) [5](#)
- [31] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023) [3](#)
- [32] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) [1](#), [3](#)
- [33] Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016) [5](#), [6](#)
- [34] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022) [3](#)
- [35] Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023) [1](#), [3](#)
- [36] Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., Shan, Y.: Gpt4tools: Teaching large language model to use tools via self-instruction. arXiv preprint arXiv:2305.18752 (2023)
- [37] Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023) [1](#), [3](#)
- [38] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022) [3](#)
- [39] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022) [6](#), [7](#)
- [40] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022) [3](#)
- [41] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) [3](#)

A Appendix

A.1 More Interactive Cases

In this section, we provide additional dialogue records of ChatSpot in this section. As shown in Figure 6 and Figure 7. ChatSpot supports multiple levels of interaction, including full image, region boxes, and region points. In the future, we will support more diverse interaction forms.

A.2 Failure Cases Analysis

Due to the limitation of data and instructions, ChatSpot may encounter challenges in certain landmark scenarios. We have conducted a comprehensive analysis and are determined to address these limitations in the next version. Specifically, as shown in Figure 8 (a), due to the limited inclusion of license plate recognition data in our dataset MGVLID, ChatSpot’s performance in license plate recognition is not satisfactory. And, as illustrated in Figure 8 (b), ChatSpot may encounter challenges in accurately identifying the precise colors of regions or objects when confronted with complex backgrounds. To address these failure cases, we can overcome them by incorporating specific instructions. In future versions, we aim to develop a wider range of instructions while also ensuring a balanced dataset.

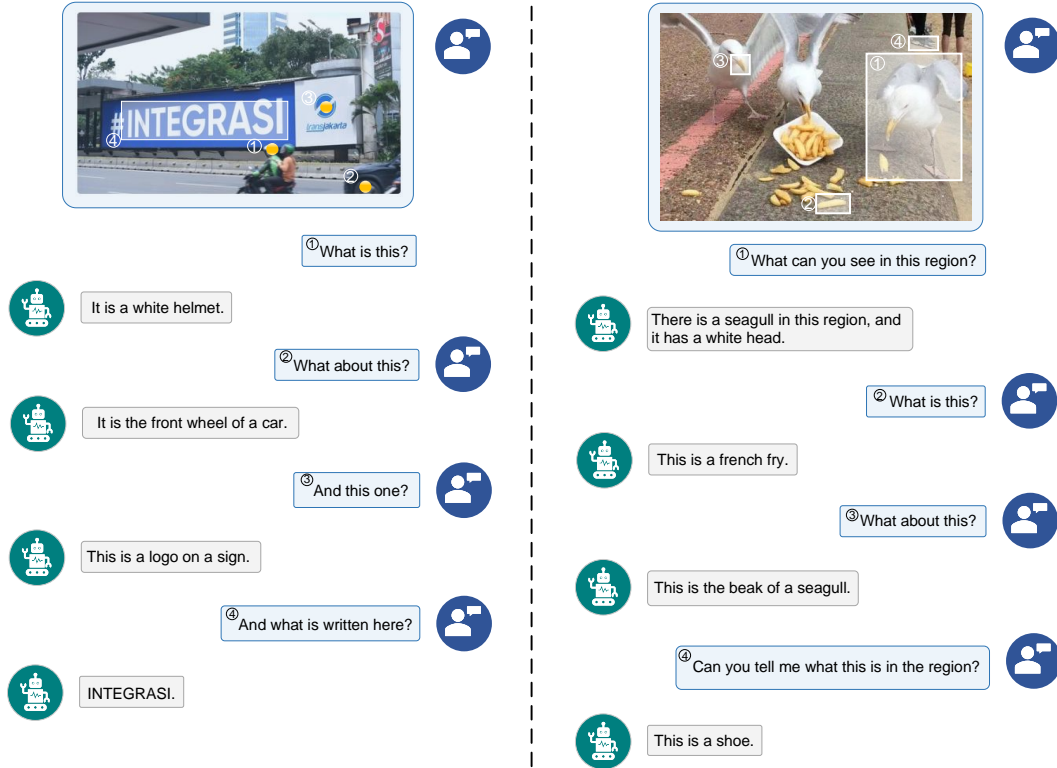
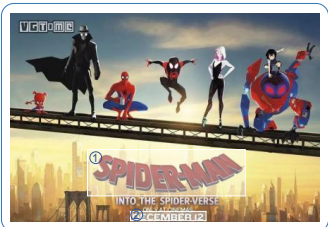


Figure 6: More cases of interacting with ChatSpot.



Which style does this movie poster belong to?


This movie poster belongs to the cartoon style.

① Can you tell me the name of this movie?

Yes, the name of this movie is "Spider-Man".

② And when can I go to the cinema to watch this movie?

You can go to the cinema on December 12th to watch this movie.



① What can you see in this area?

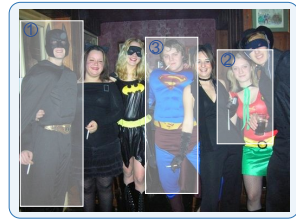
I see a small blue cup on the table.

② What is this then?

This is a small ceramic doughnut on the table.

③ Have you heard about the mathematical joke that a topologist could not distinguish a coffee mug and a doughnut?

Yes, I have heard that joke before. It is based on the fact that coffee mugs and doughnuts are topologically equivalent, meaning that they have the same shape but may not be identical in other properties.



① What role is this guy?

He is Batman, a superhero from the DC Comics universe.

② What about this guy?

He is Robin, the sidekick of Batman.

What is the relationship between these two guys in the movie?

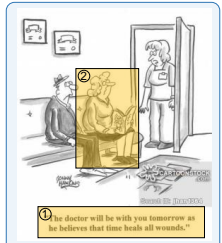
In the movie, Batman and Robin are partners, fighting crime together as a team.

③ By the way, what role is this man playing?

He is Superman, another superhero from the DC Comics universe.

What about the relationship between this guy and Batman in the movie?

In the movie, Superman and Batman are also partners, forming the Justice League to fight crime and save the world.



① What is written here?

"The doctor will be with you tomorrow as he believes that time heals all wounds."

② What is this woman doing?

She is reading a newspaper.

Can you tell me how to draw a character like this?

Certainly! To draw a character like this, you can start by sketching out the basic shape of the figure. Then, add details such as the facial features, clothing, and alignment of the features to create a balanced and visually appealing character.

Figure 7: More cases of interacting with ChatSpot.



① What is the license plate number of this car?



The license plate number of this car is G302 BFO.

(a)



① What is the color of this?



This is a black and white hat.

(b)

Figure 8: Some failure cases of ChatSpot. (a) ChatSpot on license plate recognition. (b) ChatSpot on color recognition.