Automated Inspection Report Generation Using Multimodal Large Language Models and Set-of-Mark Prompting

Hongxu Pu[1], Xincong Yang[1, 2, *], Zhongqi Shi[3, 4] and Nan Jin[3, 4]

[1] School of Civil and Environmental Engineering, Harbin Institute of Technology (Shenzhen), China
[2] Guangdong Provincial Key Laboratory of Intelligent and Resilient Structures for Civil Engineering, China
[3] Key Laboratory of Urban Safety Risk Monitoring and Early Warning, Ministry of Emergency Management, China
[4] Shenzhen Technology Institute of Urban Public Safety, China
perrykelly779@gmail.com, yangxincong@hit.edu.cn, shizq@szsti.org, jinnan@szsti.org

Abstract

In the context of the increasing expansion and complexity of civil engineering projects, construction inspection plays a crucial role in ensuring project quality and safety. The traditional construction inspection report writing process mainly relies on the manual records of on-site inspectors. This process is not only time-consuming but also easily affected by personal subjective judgments. In the current rapidly evolving construction environment, there are obvious limitations to this traditional method, especially in terms of the accuracy and timeliness of the reports. In view of this, this study proposes an innovative approach that combines the Set-of-Mark (SoM) prompting technology and the multimodal Large Language Models (LLMs), aiming to automate the construction inspection report generation process and improve the efficient and effectiveness of the onsite inspection. The case study shows that the method can fulfill the basic requirements of construction inspection reports and further improves the quality of the report in complicated scene through SoM prompting. The core of this method is to conduct a more accurate analysis of the conditions of the construction site by overlaying marks on key areas of the construction inspection images and using the multimodal LLMs to capture the region of interest (ROI), and then automatically generate detailed construction inspection reports. This technological innovation not only significantly improves the efficiency of construction inspection report writing, but also greatly enhances the quality and credibility of the report content through in-depth image analysis and text generation.

Keywords

Construction inspection, multimodal large language model, Set-of-Mark prompting, automated report generation

1 Introduction

Construction inspection work is a key component to the success of civil engineering projects, especially in today's increasingly large-scale projects and complex construction environments. To obtain a comprehensive understanding of the construction sites, engineers have to monitor the entire construction procedures and associated factors all the time, which is a process often hindered by subjectivity and inefficiency. To address this issue, this study proposes an innovative approach that leverages the Set-of-Mark (SoM) prompting in conjunction with multimodal LLMs to automate the generation of construction inspection reports.

The proposed approach entails processing construction site images using SoM prompting [1], a visual prompt technique that segments images into distinct, semantically meaningful regions, each marked with identifiable symbols. These enhanced images are then delivered to multimodal LLMs, such as GPT-4 [2], which are adept at interpreting complex visual and textual data. This collaboration has resulted in the generation of construction inspection reports that are not only detailed and accurate, but also largely avoid the problem of subjective bias that tends to occur when reports are prepared manually.

The novel approach of combining SoM prompting with multimodal LLMs in construction inspection represents a significant advancement over traditional methods. It utilizes the cutting-edge capabilities of image processing technology and language models to provide high quality construction inspection reports. This method promises to enhance the overall efficiency of construction management and quality control processes.

The subsequent sections of the paper will elaborate on the methodology of applying SoM prompting for inspection image processing, the integration of these images with multimodal LLMs, and the impact of this integration on the quality and reliability of construction inspection reports. We will also present a case study that demonstrates this method's application in a real-world engineering project and provide an analysis of the results. The conclusion will summarize the findings and explore potential future research directions.

2 Methodology

This section describes in detail the two core technologies used in our study: the SoM prompting and the multimodal LLMs. First, we introduce the SoM prompting to mark the construction inspection images, and then utilize the multimodal LLMs to convert images and texts to automatically generate construction inspection report.

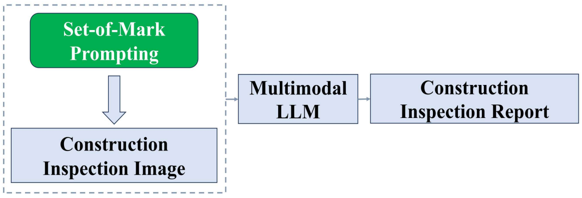

Figure 1. The workflow of the proposed approach

2.1 SoM Prompting

SoM prompting is a novel visual prompting method designed for multimodal LLMs. Its main purpose is to enhance the visual localization ability of the model through visual marks. The mathematical formulation of SoM prompting can be expressed as follows:

$$T_o = F(SoM(I), T_i) \quad (1)$$

where I is the input image, T_i is the textual query, and T_o is the textual output of the model. The key to the SoM prompting strategy is the ability to divide an image into a series of semantically meaningful regions and to impose auxiliary information on these regions to make them both interpretable and describable by multimodal LLMs.

2.1.1 Image Segmentation

Image segmentation is the first step in SoM prompting, aiming at automatically or semi-automatically extracting semantically aligned regions. For this purpose, a range of image segmentation tools have been adopted, such as SEEM [3], Semantic-SAM [4], and SAM [5]. These tools need to have the following properties, firstly robustness to ensure that the tools can segment regions with precision and convey fine-grained spatial layout information to the multimodal LLMs, secondly an open vocabulary to recognize objects outside of the predefined vocabulary, and lastly richness of granularity to support not only the segmentation of complete objects, but also the segmentation of part of an object's region.

In this study, these image segmentation tools can be used to delineate different areas of the construction site such as construction equipment, worker locations, and building materials. As shown in Figure 2, a construction inspection image I of size $A \times B$ is divided into N regions $R = [r_1, r_2, \cdots r_N] \in \{0, 1\}^{N \times A \times B}$. Each region can be represented by a binary mask, where each mask corresponds to a key region in the image. For example, one mask may represent the location of an air conditioning duct, while another mask may represent the area where construction materials are stacked.

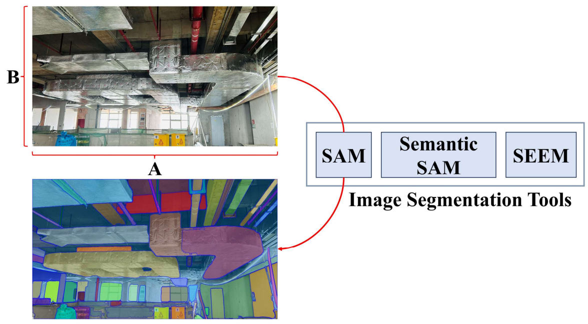

Figure 2. Image segmentation using SAM

2.1.2 Mark Generation

After completing the image segmentation, marks that contribute to multimodal LLM localization need to be generated for each region. The type of marks depends on whether they can be interpreted by the multimodal LLMs. In this study, considering the complexity of the construction site scene, alphanumeric marks should be used because they not only take up less image space, but also can be recognized by most multimodal LLMs using their OCR function [6]. In addition to this, the polygon and box marks used in image segmentation will be used as auxiliary marks.

In addition to selected mark types, assigning optimally located marks to each significant region in the inspection image is critical for the generation of subsequent inspection report. There is a significant problem in current mark assignment methods: placing the marks directly at the center of the mask leads to overlapping between marks. To solve this problem, an optimization algorithm is employed as shown in Figure 3, which are based on the principle of reducing the overlap between marks by calculating the area of the identified regions in the image and sorting them in order of size to ensure that smaller regions are processed first before larger ones when assigning marks. The algorithm further excludes overlaps with processed regions to obtain an independent mask, and then applies a distance transformation algorithm to find the optimal mark locations.

```
def   Find_Optimal_Mark_Location (region):
      optimal_location = calculate_optimal_location (region)
      return optimal_location

def   Assign_Marks_Optimally (regions):
      sorted_regions = sort_by_area (regions)
      mark_locations = {}

      for region in sorted_regions:
          if not is_overlap (region, processed_regions):
              mark_location = Find_Optimal_Mark_Location (region)
              processed_regions.append (region)
              mark_locations [region] = mark_location

      return mark_locations
```

Figure 3. Pseudocode of mark allocation algorithm

By using the image segmentation tools and the mark allocation algorithm, the inspection image processed by SoM prompting can be obtained as shown in Figure 4.

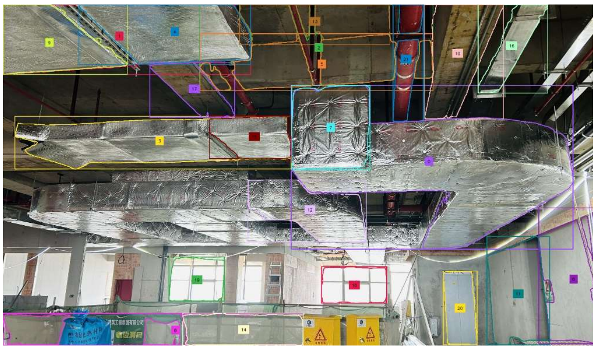

Figure 4. Inspection image with SoM prompting

2.2 Multimodal LLMs

Multimodal LLMs represent an important advancement in the field of artificial intelligence (AI), and their core capability is to concurrently process and understand data from different modalities, such as images, text, and sound [7]. These models are typically based on deep learning architectures, in particular neural networks such as convolutional neural networks (CNNs) [8] and vision transformer (ViT) [9] for image processing, and some transformer models [10] for processing sequential data such as text.

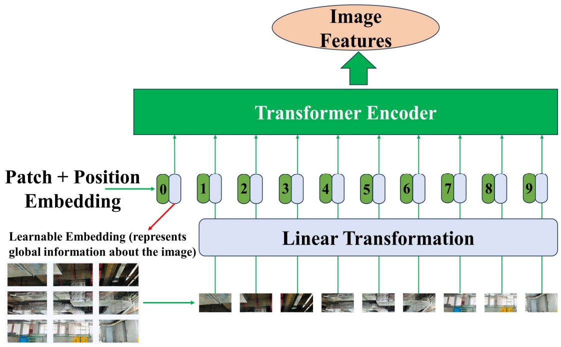

Figure 5. The process of extracting image features by vision transformer

2.2.1 Selection of Multimodal LLMs

In the process of automatically generating construction inspection reports, multimodal LLMs not only extract information from inspection images, but also text them to generate inspection reports. Therefore, the selection of an appropriate multimodal LLM is an important factor in the quality of the report. The choice of model relies on a comprehensive evaluation of the model in multiple aspects [11]. First, the model must demonstrate strong perception ability, which includes accurately identifying specific objects and details on the construction site. For example, the model needs to be able to identify construction equipment, engineering materials, and even safety signs during construction. Furthermore, cognitive capability plays a crucial role. The model is expected to interpret visual data effectively and translate it into a logically coherent and comprehensive construction inspection report. This requires that the model is not just a simple information processing tool but has certain reasoning and logical analysis capabilities.

Secondly, the model should have good instruction following ability. In the process of generating construction inspection reports, the model needs to perform tasks according to specific instructions, such as generating specific project status descriptions based on construction images. Therefore, the selected multimodal LLM should be able to accurately understand and respond to these instructions to avoid misunderstandings or biases.

In addition, preventing hallucination problems [12] is also an important criterion for selecting multimodal LLMs. When dealing with complex construction scenarios, the model should not produce incorrect object recognition, such as incorrectly identifying engineering hazards or equipment that do not exist. Therefore, models with higher accuracy and reliability are more appropriate.

In practical applications, considering the particularity of construction inspections, the training data and instruction design of the model are also very critical. The training data of the model should cover a wide range of construction scenarios to ensure its effectiveness and accuracy in practical applications. Meanwhile, the instruction design should be as concise and clear as possible so that the model can accurately interpret and execute it.

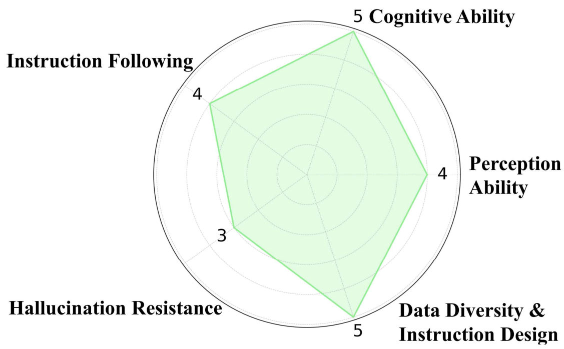

Figure 6. The criteria and weights of model selection

Taking the above criteria into consideration, when selecting a multimodal LLM suitable for construction inspection scenarios, it is recommended to choose those models that perform well in perception and cognitive tests, can accurately follow instructions, and have good performance in object recognition and logical reasoning. In addition, considering the complexity and variability of construction inspections, it will be more advantageous to choose a model that can quickly adapt to new scenarios and tasks. Through careful selection, multimodal LLMs can effectively improve the quality of automated generation of construction inspection reports, thereby improving overall project management efficiency and safety.

2.2.2 Construction Inspection Report Generation

In multimodal LLMs, the image and text inputs are first processed through their respective preprocessing modules and then embedded into the same dimensional space. For image data, this may include steps such as feature extraction, object recognition, and image segmentation, which are designed to extract useful visual information from the image. For text data, word embedding is a key step. Word embedding converts each word into a vector in a high-dimensional space, and these vectors are usually learned by training algorithms such as Word2Vec [13] on large text datasets. Mathematically, this can be expressed as a mapping function as follows:

$$f\colon word \to \mathbb{R}^d \tag{2}$$

where $word$ is a word in the vocabulary and $\mathbb{R}^d$ is a d-dimensional vector space. The vector representation of each word not only captures the semantic information of the word, but semantically similar words are mapped to locations close to each other in this high-dimensional space.

The processed data is then fed into the core of the model, which typically contains multiple levels of network structure for in-depth analysis and fusion of information from different modalities. In this process, the model learns how to correlate and interpret data from different sources. Considering that the data is different for each modality, the fusion process needs to combine this information efficiently. For example, if there are an image feature vector $v_i \in \mathbb{R}^m$ and a text feature vector $v_t \in \mathbb{R}^n$, a common fusion strategy is to use weighted sum [14] or concatenation [15]. The mathematical principle of weighted sum is as follows:

$$v = \alpha v_i + (1-\alpha) v_t \tag{3}$$

where α is a weight parameter to balance the importance of the two models. Another common approach concatenation is to directly concatenate the two types of feature vectors to form a new vector:

$$v = [v_i; v_t] \tag{4}$$

where $[.;.]$ represents the concatenation of vectors. In both approaches, the key is to select and tune the parameters to combine information most effectively from different sources. In multimodal learning, this is usually achieved by optimizing an objective function that quantifies the degree of match between the fused data and the expected output. With such a fusion approach, the model can synthesize information from different modalities and improve its ability to understand complex data.

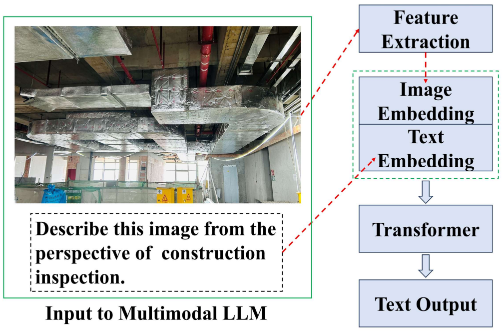

Figure 7. The flowchart of construction inspection report generation

On the output side of a multimodal LLMs, the model can generate a comprehensive output based on the input data, which may be a report that combines image and textual information, or a comprehensive interpretation and analysis of the input data. This capability makes the multimodal LLMs particularly suitable for complex tasks that require the concurrently understanding of visual and textual information, such as the automated generation of construction inspection reports.

In the construction inspection scenario, the multimodal LLMs can use its ability to understand images to identify specific features of the construction site, while using its text processing capabilities to generate a detailed report. This combination of visual and textual analysis makes multimodal LLMs ideal for generating accurate and comprehensive construction inspection reports.

3 Case Study

This case study focused on a large-scale construction project under construction in Pingshan District, Shenzhen City, Guangdong Province. This project provides a valuable practical scenario for in-depth analysis and verification of the application of the multimodal LLMs in the automated generation of construction inspection reports. The project covers an area of 0.2 square kilometers, providing diverse construction activities and rich visual data for multimodal LLMs. Since the MiniGPT-4 multimodal LLM [16] has demonstrated excellent performance in several benchmark tests, we chose MiniGPT-4 for the case study.

MiniGPT-4 is designed to maintain the original efficient processing power of GPT-4 while optimizing its size and resource consumption to make it more suitable for application scenarios that require faster speeds and fewer computational resources. It is specifically designed for multimodal tasks, including image and text processing. It utilizes the state-of-the-art LLM Vicuna [17] as a decoder, combined with the pre-trained visual component BLIP-2 containing ViT and Q-Former [18]. MiniGPT-4 aligns the encoded visual features with the Vicuna language model through a single projection layer and fixes the other visual and language components. The initial training of the model covers a large image-text alignment dataset, which is then fine-tuned with high-quality, accurately aligned image-text pairs in a second phase, a step that is critical to improving the reliability and overall usability of the model. MiniGPT-4 demonstrates a variety of GPT-4-like capabilities, including detailed image description generation and website creation based on handwritten text instructions.

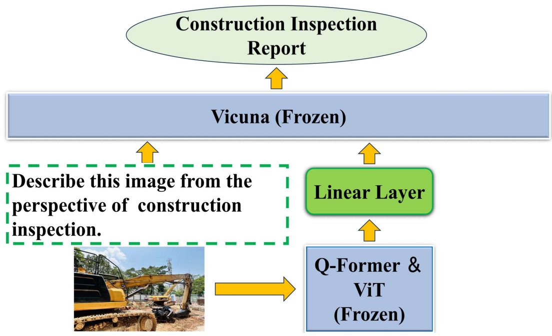

Figure 8. The architecture of MiniGPT-4

To meet the computational requirements of the project, a high-performance desktop computer with an RTX 4080 16GB was used to ensure the efficiency and stability of the process of using the MiniGPT-4 to generate construction inspection reports.

The case study first explored the ability of the MiniGPT-4 model to rely on original inspection image which is not combined with SoM prompting to generate construction inspection report.

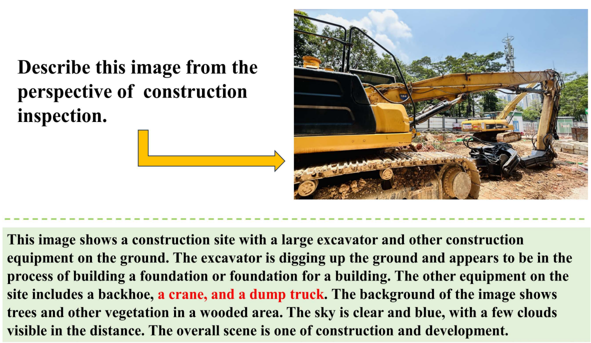

Figure 9. Construction inspection report generated by MiniGPT-4 for inspection image

without SoM prompting

By analyzing the inspection report shown in Figure 9, it can be found that the inspection report comprehensively and accurately describes the contents of the construction inspection image. Although there are errors in the identification of objects such as crane and dump truck marked in red in the report, the generated report still adequately fulfills the standards required for construction inspection report.

Further case study introduced SoM prompting, which aimed to enhance the model's ability to identify and describe construction details by annotating key areas of construction inspection images.

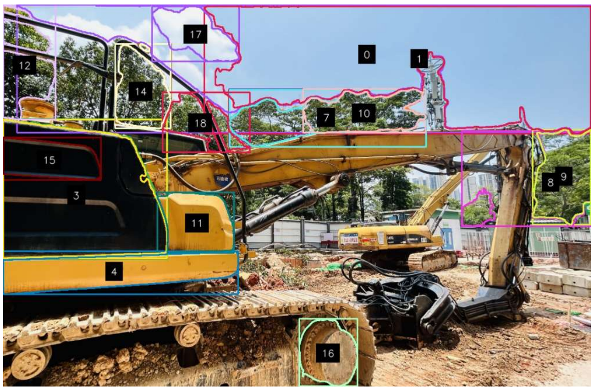

Figure 10. Construction inspection image marked by SoM prompting

With the assistance of SoM prompting technology, the MiniGPT-4 processed the annotated images and generated detailed construction inspection report. The inspection report generated by MiniGPT-4 for the image annotated by SoM prompting is shown in Figure 11.

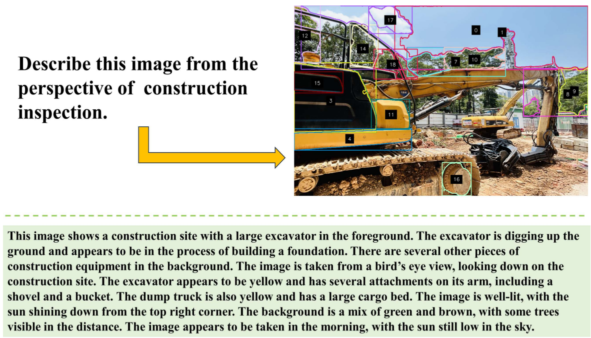

Figure 11. Construction inspection report generated by MiniGPT-4 for inspection image with SoM prompting

The results shown that although the MiniGPT-4 can generate basic construction inspection reports by analyzing construction inspection images that have not been processed by SoM prompting, the introduction of SoM prompting has significantly improved the detail and accuracy of the report, which verifies the value of the multimodal LLMs and SoM prompting for the application in real construction scenarios.

4 Conclusion and Future Work

In this study, we successfully introduced the MiniGPT-4 to a large-scale construction project in Shenzhen and realized the automated generation of construction inspection reports that obtain a comprehensive understanding of the field scene. This result demonstrates the potential of multimodal LLMs in the construction industry. Multimodal LLMs can effectively analyse complex visual data at construction sites and successfully capture key factors of construction activities, thereby generating comprehensive and accurate construction inspection reports. Furthermore, we also improved the accuracy of reports by introducing SoM prompting, effectively improving the efficiency and quality of construction scene data analysis.

However, this method also has certain limitations. Due to the huge number of parameters of the multimodal LLMs, the automated generation of construction inspection reports places high demands on computing resources. The running and processing of the current multimodal LLMs requires powerful hardware support, which may limit its feasibility in some practical application scenarios. In addition, the effectiveness of the report generated by the multimodal LLMs depends on the quality and diversity of the training data used to pre-train the models, which also places higher requirements on the source of training data for the multimodal LLMs.

In the future, we aim to harness the specialized knowledge within the construction field to fine-tune pre-trained multimodal LLMs, and by leveraging cloud computing resources, we can distribute the computing load and reduce the time and local resource requirements. Additionally, the exploration of more lightweight LLMs architectures would be beneficial. These architectures can remain the essential features necessary for accurate report generation while minimizing the model size and complexity. Moreover, we also plan to further develop tools and user-friendly interfaces for real-time analysis of construction inspection images so that construction managers can more easily use and understand the reports generated by the model. Through these measures, we look forward to advancing the automation process within the construction industry and providing new ideas and solutions for the practical application of multimodal LLMs in the field of civil engineering. As technology continues to develop, models like MiniGPT-4 will play an increasingly important role in the digital transformation of the construction industry.

5 Acknowledgements

This research is supported by Key Technologies R&D Program (2022YFC3801203), supported by National Natural Science Foundation of China (Grant No. 52108286), supported by Shenzhen Science and Technology Programs (Grant No. GXWD20220818002513001, Grant No. RCBS20221008093128076).

References

[1] Yang J, Zhang H, Li F, Zou X, Li C, Gao J. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V [J]. arXiv preprint arXiv:2310.11441, 2023.

[2] OpenAI. GPT-4 Technical Report[J]. arXiv preprint arXiv:2303.08774, 2023.

[3] Zou X, Yang J, Zhang H, et al. Segment everything everywhere all at once[J]. arXiv preprint arXiv:2304.06718, 2023.

[4] Li F, Zhang H, Sun P, et al. Semantic-sam: Segment and recognize anything at any granularity[J]. arXiv preprint arXiv:2307.04767, 2023.

[5] Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[6] Liu Y, Li Z, Li H, et al. On the hidden mystery of ocr in large multimodal models[J]. arXiv preprint arXiv:2305.07895, 2023.

[7] Yin S, Fu C, Zhao S, et al. A Survey on Multimodal Large Language Models[J]. arXiv preprint arXiv:2306.13549, 2023.

[8] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern recognition, 2018, 77: 354-377.

[9] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[11] Fu C, Chen P, Shen Y, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models[J]. arXiv preprint arXiv:2306.13394, 2023.

[12] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[13] Church K W. Word2Vec[J]. Natural Language Engineering, 2017, 23(1): 155-162.

[14] Marler R T, Arora J S. The weighted sum method for multi-objective optimization: new insights[J]. Structural and multidisciplinary optimization, 2010, 41: 853-862.

[15] Zhang C, Yang Z, He X, et al. Multimodal intelligence: Representation learning, information fusion, and applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478-493.

[16] Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.

[17] Chiang W L, Li Z, Lin Z, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality[J]. See https://vicuna. lmsys. org (accessed 14 April 2023), 2023.

[18] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[J]. arXiv preprint arXiv:2301.12597, 2023.