SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2026, Vol. 69, Iss. 2, 122103:1–122103:17
https://doi.org/10.1007/s11432-024-4416-8

PGPL: enhancing spatial awareness abilities of multimodal large language models based on precise geometric position learning

Yongqiang ZHAO[1,2†], Zhenyu LI[3†], Zhi JIN[1,2*], Feng ZHANG[3*], Ziliang WANG[1,2], Lianwei WU[4], Chengfeng DOU[1,2], Haiyan ZHAO[1,2] & Xinhai XU[3*]

[1]*Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing 100871, China*
[2]*School of Computer Science, Peking University, Beijing 100871, China*
[3]*Academy of Military Sciences, Beijing 100089, China*
[4]*School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*

Received 30 July 2024/Revised 13 February 2025/Accepted 24 April 2025/Published online 9 October 2025

Abstract Multimodal large language models (MLLMs) have already begun to be used in visual question answering (VQA), autonomous driving, and smart healthcare, showing great application potential. However, existing MLLMs have significant gaps compared with human intelligence in terms of spatial awareness tasks, especially in accurately identifying and interpreting complex spatial relationships between target entities. This deficiency severely impacts the accuracy of VQA, the safety of autonomous driving, and the reliability of smart healthcare. In order to meet the requirements for the accuracy of spatial relationship recognition in specific applications, we propose a novel framework named PGPL, which attempts to enhance the spatial awareness ability of an MLLM by integrating precise geometric position information between target entities on the MLLM without the need for additional training of the MLLM. Specifically, the PGPL framework leverages the spatial position generation model and the scene graph generation model to obtain geometric absolute position and geometric relative position of the target entities in the visual input. And then, it introduces a multidimensional information fusion strategy to guide the MLLM to accurately answer questions related to spatial awareness. The quantitative experimental results of six popular datasets and twelve MLLMs, as well as the related qualitative experimental results, fully demonstrate the importance of the precise geometric position information for correctly answering spatial awareness questions, and demonstrate the superiority of the PGPL framework.

Keywords multimodal large language model, spatial awareness, geometric position learning, geometric absolute position, geometric relative position, multidimensional information fusion

Citation Zhao Y Q, Li Z Y, Jin Z, et al. PGPL: enhancing spatial awareness abilities of multimodal large language models based on precise geometric position learning. Sci China Inf Sci, 2026, 69(2): 122103, https://doi.org/10.1007/s11432-024-4416-8

1 Introduction

Multimodal large language models (MLLMs) [1–4] have become a focal point of research. They harness powerful large language models (LLMs) [5–9] as the cognitive engines and possess the abilities in understanding and interpreting multimodal data. This showcases their vast potential across diverse applications [10–15] such as visual question answering (VQA), autonomous driving, smart healthcare, intelligent manufacturing, general-purpose robots, and virtual reality.

However, among these applications, there are many scenarios that require very powerful spatial awareness abilities. For instance, in VQA [16–18], accurately positioning and identifying spatial relationships between objects are critical for answering user queries. Especially in autonomous driving [19,20], precise localization of objects such as vehicles, pedestrians, traffic signs, road markings, and parking spaces is crucial to ensure safe driving and prevent accidents. Smart healthcare [21,22] also needs to accurately localize the pathological areas and identify the spatial relationships between anatomical structures for enhancing diagnostic accuracy, formulating effective treatment plans, and guiding surgical procedures. Despite their potential, existing MLLMs [2, 3, 23] exhibit significant limitations in capturing and interpreting complex spatial relationships. Recent studies have systematically evaluated the spatial reasoning

* Corresponding author (email: zhijin@pku.edu.cn, zhangfeng@nudt.edu.cn, xuxinhai@nudt.edu.cn)
† These authors contributed equally to this work.

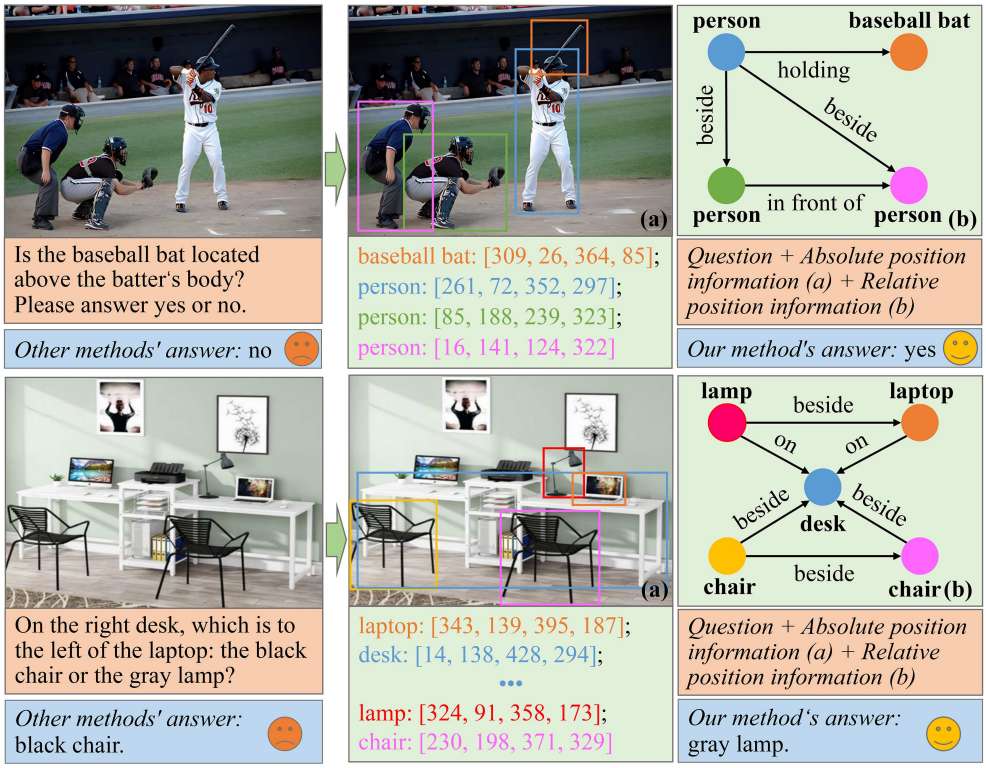

Figure 1 (Color online) Examples of existing MLLMs (left) and our proposed framework (right) in VQA.

capabilities of popular vision-language models, revealing that even the best-performing models achieve only 56% accuracy on basic spatial reasoning tasks, compared to nearly 99% human-level performance [24]. This deficiency severely impacts the accuracy of VQA, the safety of autonomous driving, and the reliability of smart healthcare. As illustrated by the two examples on the left side of Figure 1, the existing MLLMs [2,3] have significant challenges in accurately identifying the spatial relationships between target objects within the VQA application, often giving wrong answers.

Current approaches to improving MLLMs' spatial awareness heavily rely on resource-intensive pre-training or fine-tuning [3,23,25,26], which demand substantial computational resources and extensive data preparation. For humans, an effective method to solve the aforementioned problem is to use precision measuring tools to obtain precise spatial position information. Inspired by human intelligence, which leverages precision tools for spatial reasoning, we propose PGPL (precise geometric position learning), a novel framework that enhances the spatial awareness of MLLMs by integrating absolute and relative geometric positional information without requiring additional data or retraining. Specifically, the PGPL framework employs the spatial position generation model to identify the categories and corresponding geometric absolute positional coordinates of entities in the visual input. Simultaneously, it utilizes the scene graph generation model to extract the geometric relative positional relationships among entities. These two types of positional information—absolute and relative—are seamlessly integrated through a multidimensional information fusion strategy, enhancing the ability of MLLMs to reason more effectively about spatial relationships. The examples on the right in Figure 1 demonstrate our framework's accurate identification of object spatial relationships.

Our contributions can be summarized as follows.

• We propose the novel PGPL framework, designed to enhance the spatial awareness abilities of MLLMs. This framework leverages spatial positional information to provide auxiliary information about absolute geometric positions and utilizes scene graph information to supply auxiliary insights into relative geometric positions. It also incorporates a multidimensional fusion strategy to guide MLLMs to generate more accurate results, offering a comprehensive solution ranging from basic spatial positioning to understanding complex object relationships within a scene.

• We systematically investigate the impact of various pre-trained models on overall performance and concurrently clarify the criteria for selecting spatial position generation models, precisely identifying the optimal selection of corresponding parameters.

• We conduct extensive experiments on six popular datasets and twelve base models. The experimental results robustly demonstrate that our framework significantly enhances the performance of MLLM in spatial awareness and related task aspects.

2 Related work

Multimodal large language model (MLLM). The MLLM [4,27–29] is capable of understanding and processing diverse information from different modalities, playing a crucial role in advancing multimodal research and application. The MLLM is divided into two main types. The first type involves combinations of LLM and visual encoders aiming to achieve multimodal understanding at minimal training costs. Examples include LLaVA [2], BLIP [3], mPLUG-Owl [30], GPT4RoI [31], and Kosmos [32]. The second type involves the integration of LLM with various smaller multimodal models. These models use the LLM as a central hub, invoking different small models to perform various multimodal tasks, and ultimately consolidating into user responses. Examples include HuggingGPT [33], CompeGPT [34], and Visual ChatGPT [35]. Our study, however, focuses its research on the first type of MLLM.

Spatial awareness multimodal large language models. The current methods to enhance the spatial perception capabilities of multimodal large language models can be divided into two main categories. The first category involves constructing large-scale real image-text pair datasets to pre-train or fine-tune baseline multimodal large language models [32,36]. These datasets include images and text pairs with precise location information, enabling the models to better learn fine-grained spatial relationships and localization capabilities. The second category focuses on making the models learn the specific spatial location information corresponding to the text during the pre-training or fine-tuning process, thereby enhancing their spatial perception abilities [31,37]. Both methods incur significant costs in terms of data construction and computational resources. To address these challenges, we propose the PGPL framework, which aims to bolster the spatial awareness abilities of MLLMs without requiring additional training of the MLLM.

Spatial position generation (SPG). Spatial position generation models refer to a set of models that can extract an entity's geometric absolute spatial position information from visual inputs. These models excel at extracting detailed geometric information, including the boundaries, coordinates, and depth of objects within a visual scene. The suite of models in this domain is diverse, extending to but not restricted to, object detection models [38–40], image segmentation models [41–43], video segmentation models [44,45], and depth estimation models [46,47].

Scene graph generation (SGG). Scene graph generation models [48–51] are designed to automatically convert an image into a structured semantic graph. These models identify and label objects within the image, discern the spatial position relationships between these objects, and organize this information into a graph structure where nodes represent the objects and edges denote the relationships. These models can accurately capture the semantic information of complex scenes in images, including the geometric relative positional relationships between entities.

Visual question answering, referring expression comprehension, and image captioning. This paper uses the VQA application [52–54] as an example to specifically analyze how to enhance the spatial awareness ability of MLLM. VQA involves analyzing images to answer text questions related to their content. To provide relevant answers, the VQA model requires a deep understanding of the image content and the question text, encompassing awareness of objects, scenes, and spatial relationships within images, as well as comprehension of natural language questions. To more broadly validate the effectiveness of our PGPL framework, we also conducted performance validation on referring expression comprehension and image captioning applications. Referring expression comprehension (REC) [55] involves identifying and locating specific objects mentioned in the text within an image based on the text description. Image captioning [56] involves using a computer to automatically generate a complete, smooth, and suitable caption for a given image, thereby realizing the multimodal conversion from image to text.

3 Method

In this section, we first provide an overview of the proposed framework, PGPL, followed by a detailed description of its core components: target entity extraction, geometric absolute position learning, geometric relative position learning, and multidimensional information fusion.

3.1 Overview

Existing studies [3, 23, 25, 26] suggest that enhancing the MLLM's capabilities through targeted training methods incurs a high cost in terms of both data construction and computational resources. Therefore, we propose the

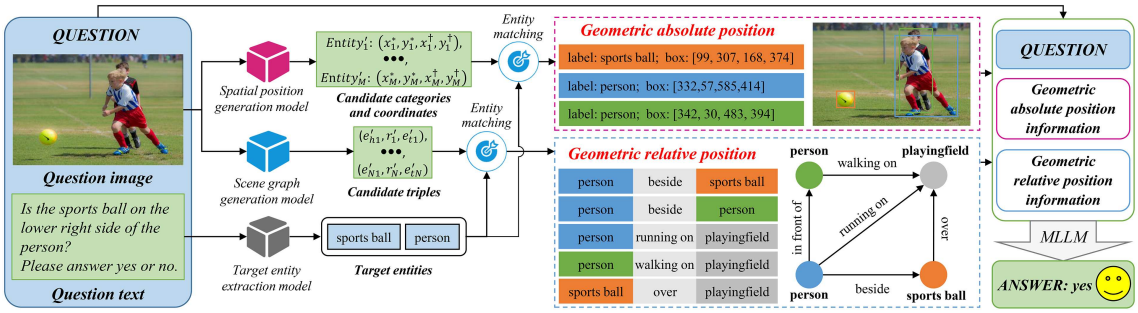

Figure 2 (Color online) Overview of our PGPL framework in VQA. The input QUESTION consists of two components: question image and question text. The target entity extraction model processes the question text to obtain the target entities. The spatial position generation model processes the question image to obtain corresponding candidate categories and coordinates, and matches them with the target entities to obtain geometric absolute position. The scene graph generation model processes the question image to obtain corresponding candidate triplets, and matches them with the target entities to obtain geometric relative position. The MLLM integrates the initial QUESTION, geometric absolute position information and geometric relative position information of the target entities to obtain the final output ANSWER.

PGPL framework to bolster the spatial awareness abilities of MLLMs:

$$\text{ANSWER} = \text{PGPL}(\text{QUESTION}), \tag{1}$$

where QUESTION is the user's input request, and ANSWER is the response obtained through the PGPL framework. The framework attempts to enhance the spatial awareness ability of an MLLM by integrating precise geometric position information between target entities without the need for additional training of the MLLM.

The PGPL framework mainly includes four parts: target entity extraction (Part 1: Subsection 3.2), geometric absolute position learning (Part 2: Subsection 3.3), geometric relative position learning (Part 3: Subsection 3.4), and multidimensional information fusion (Part 4: Subsection 3.5). As shown in Figure 2, for a given multimodal QUESTION, the PGPL framework utilizes pre-trained models to acquire geometric absolute position and geometric relative position of entities pertinent to spatial awareness questions. It then guides the MLLM to respond to QUESTION based on this information, thereby enhancing the MLLM's abilities in spatial awareness tasks. In detail, the framework first utilizes the target entity extraction model to extract the target entities involved in determining spatial relationships from the Q_T in the multimodal QUESTION (Part 1). Subsequently, it employs the spatial position generation model and scene graph generation model to obtain the spatial positional information (candidate categories and coordinates) and scene graph information (candidate triplets) of entities within the Q_I of the QUESTION, and then uses the entity matching model to extract the geometric absolute position and geometric relative position of the target entities (Parts 2 and 3). Finally, the PGPL framework guides the MLLM to leverage the geometric absolute position information and geometric relative position information of the target entities to address spatial awareness-related QUESTION, resulting in the ultimate ANSWER (Part 4). It should be noted that the MLLM within the PGPL framework can be replaced as needed.

3.2 Target entity extraction

In VQA, the QUESTION primarily comprises two components: the user-input question text data (Q_T) and multimodal visual data (Q_I). This paper employs the target entity extraction (TEE) model to extract the target entities among which the spatial relationships are required by the user-input Q_T, as demonstrated in

$$(E_1, \ldots, E_k) = (\text{TEE})(Q_T), \tag{2}$$

where (E_1, …, E_k) represents the target entities requiring spatial relationship determination within the Q_T, with k indicating the number of target entities. (E_1, …, E_k) represents the target entities requiring spatial relationship determination within the Q_T, with k indicating the number of target entities.

The TEE model in this paper uses the pre-trained BERT model [57] for named entity recognition and combines it with a rule-based method. The reason for choosing the BERT model is its bidirectionality and strong context understanding ability. The BERT model is fine-tuned on the CoNLL-03 dataset [58] and is suitable for named entity recognition tasks. After all the named entities in the Q_T are identified, the target entity is determined using the rule-based method. Specifically, entities that contain relational words (such as "on the lower right side" in Figure 2) within the Q_T are identified as target entities (such as "sports ball" and "person" in Figure 2). Notably, when no relational words exist between entities in the Q_T, all named entities are considered target entities.

3.3 Geometric absolute position learning

Geometric absolute position information refers to the geometric absolute coordinate details among entities within visual inputs. This information can be obtained through various spatial position generation models, such as object detection models [59, 60], image segmentation models [61, 62], video segmentation models [63, 64], and depth estimation models [65, 66]. This subsection provides a detailed explanation of the methodology using object detection models as an example. The subsequent experimental process thoroughly investigates the effects of different types of models (Subsection 4.5).

To be more specific, we initially utilize a SPG model [67] to obtain the candidate entities' categories (E'_i) and geometric coordinates ($x_i^*, y_i^*, x_i^\dagger, y_i^\dagger$) in the input image Q_I, as illustrated in

$$E'_1 : (x_1, y_1, x_1^\dagger, y_1^\dagger), \ldots, E'_M : (x_M, y_M, x_M^\dagger, y_M^\dagger) = \mathrm{SPG}(Q_I), \tag{3}$$

where $E'_1 : (x_1, y_1, x_1^\dagger, y_1^\dagger), \ldots, E'_M : (x_M, y_M, x_M^\dagger, y_M^\dagger)$ represents the set of detected candidate entities, each characterized by its respective category and geometric coordinates. Here, M denotes the total number of detected entities, and x_i, y_i, $x_i^\dagger$, and $y_i^\dagger$ denote the coordinates of the entity i in the corresponding space. The symbols $*$ and $\dagger$ represent the coordinates of the top-left and bottom-right corners of the bounding box in detection results, respectively. The results of the categories, geometric coordinates, and number of entities depend on the capabilities of the SPG model.

Next, the entity matching (EM) model is employed to associate the categories of candidate entities $(E'_1, \ldots, E'_M)$ extracted from the visual input with the categories of target entities $(E_1, \ldots, E_k)$. This process generates a dictionary containing the geometric absolute position information of the target entities in the input image:

$$E_1, \ldots, E_k = \mathrm{EM}((E_1, \ldots, E_k); (E'_1, \ldots, E'_M)). \tag{4}$$

Here, $E_1 : (x_1, y_1, x_1^\dagger, y_1^\dagger), \ldots, E_k : (x_k, y_k, x_k^\dagger, y_k^\dagger)$ denotes the obtained geometric absolute position information of the target entities (e.g.,"sports ball" and "person" in Figure 2).

Specifically, we utilize the pre-trained BERT model to transform the candidate entities $(E'_1, \ldots, E'_M)$ and the target entities $(E_1, \ldots, E_k)$ into their respective word vectors. Subsequently, we calculate the cosine similarity between these word vectors:

$$\text{cosine similarity} = \frac{E_{iV} \cdot E'_{jV}}{|E_{iV}| \cdot |E'_{jV}|}. \tag{5}$$

Here, E_{iV} denotes the word vector corresponding to the i-th target entity, while E'_{jV} denotes the word vector corresponding to the j-th candidate entity, cosine similarity represents the calculated cosine similarity result. When the cosine similarity exceeds a predefined threshold, the two entities are considered synonyms. In this case, the coordinate information corresponding to the candidate entity is used as the geometric absolute position information for the target entity.

3.4 Geometric relative position learning

Geometric absolute position information primarily focuses on identifying and locating the geometric absolute coordinates of entities. However, in certain application scenarios, spatial awareness requires more than just understanding the geometric absolute coordinates. It also necessitates a comprehension of higher-level geometric relative position information among entities, such as the relative position relationships between entities and between entities and the scene. Geometric absolute position information refers to the 2D coordinates between target entities, while geometric relative position information encompasses higher-level relationships such as actions, orientation, and interactions. Although absolute coordinates can imply some relative positions, many complex relationships, such as "walking on", "in front of", and "hugging", cannot be captured solely by absolute coordinates. Therefore, we propose combining geometric absolute and relative position information to form a comprehensive framework. This study utilizes scene graph generation models to obtain the geometric relative position information. By combining geometric absolute and relative position information from multimodal visual inputs, we achieve a more comprehensive and accurate understanding of images. This method not only identifies the positions of objects but also understands their interactions within the scene, which is particularly beneficial for image understanding with complex scenes.

Specifically, we begin by utilizing an SGG model [68] to generate the scene graph candidate triples of the multimodal visual input, as illustrated in

$$(e'_{h1}, r'_1, e'_{t1}), \ldots, (e'_{hN}, r'_N, e'_{tN}) = \mathrm{SGG}(Q_I), \tag{6}$$

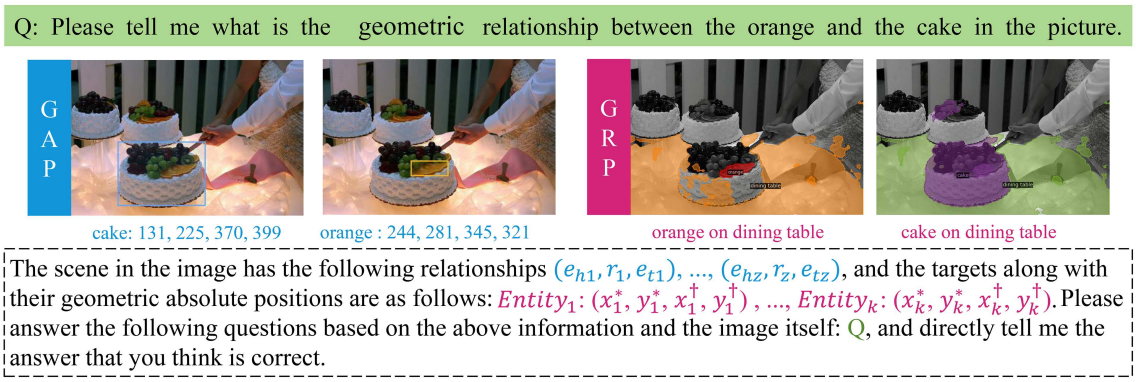

Figure 3 (Color online) Visualization of the prompt and its corresponding GAP and GRP illustrations in the PGPL framework.

where Q_I refers to the input multimodal visual data, and $\{(e'_{h1}, r'_1, e'_{t1}), \ldots, (e'_{hN}, r'_N, e'_{tN})\}$ represents the collection of candidate triples of scene graph obtained from the image. Here, e'_h represents the head entity, r' represents the relationship, and e'_t represents the tail entity, with N indicating the number of triples. Simultaneously, based on the collection of image scene graph candidate triples $\{(e'_{h1}, r'_1, e'_{t1}), \ldots, (e'_{hN}, r'_N, e'_{tN})\}$, we extract all triples relevant to the target entities mentioned in the question text Q_T using the EM model:

$$(e_{h1}, r_1, e_{t1}), \ldots, (e_{hz}, r_z, e_{tz}) = \text{EM}((E_1, \ldots, E_k); (e'_{h1}, r'_1, e'_{t1}), \ldots, (e'_{hN}, r'_N, e'_{tN})), \tag{7}$$

where $(e_{h1}, r_1, e_{t1}), \ldots, (e_{hz}, r_z, e_{tz})$ represents the resulting set of target triplets, and z indicates the total number of target triplets identified. Specifically, we utilize the pre-trained BERT model to transform the entities $(e'_{h1}, e'_{t1}, \ldots, e'_{hN}, e'_{tN})$ in the scene graph candidate triplet set $\{(e'_{h1}, r'_1, e'_{t1}), \ldots, (e'_{hN}, r'_N, e'_{tN})\}$ and the target entities $(E_1, \ldots, E_k)$ into their respective word vectors. Subsequently, we calculate the cosine similarity between these word vectors:

$$\text{cosine similarity} = \frac{E_{iV} \cdot e'_{tV}}{|E_{iV}| \cdot |e'_{tV}|}. \tag{8}$$

Here, E_{iV} denotes the word vector corresponding to the i-th target entity, while e'_{tV} denotes the word vector corresponding to the t-th entity in the scene graph candidate triplet set, cosine similarity represents the calculated cosine similarity result. When the cosine similarity exceeds a predefined threshold, the two entities are considered synonyms. The matching criterion is as follows: if there is an entity in the scene graph candidate triplet set $\{(e'_{h1}, r'_1, e'_{t1}), \ldots, (e'_{hN}, r'_N, e'_{tN})\}$ that matches with target entities $(E_1, \ldots, E_k)$ (such as "sports ball" and "person" in Figure 2), then this triplet is retained.

3.5 Multidimensional information fusion

After learning the geometric absolute and relative position information of entities related to the spatial relationship question, the critical task becomes how to effectively leverage this information to guide MLLM to accurately answer user questions related to spatial awareness. Based on the existing studies [69, 70], this study proposes a multidimensional information fusion method based on zero-shot learning, detailed in (9) and Figure 3. This zero-shot strategy uniquely enables the MLLM to utilize spatial awareness information from pre-trained models without the need for additional fine-tuning, significantly enhancing its ability to accurately interpret and respond to user questions about spatial relationships.

$$\text{ANSWER} = \text{MLLM}(\text{QUESTION},\ \text{GAP},\ \text{GRP}), \tag{9}$$

where QUESTION refers to the input comprising multimodal visual data Q_I and textual data Q_T, while GAP and GRP denote the generated geometric absolute position and geometric relative position information of the target entity, respectively.

4 Experiments

This section provides a comprehensive analysis of our proposed PGPL framework through a series of experiments. We begin by introducing the implementation details (Subsection 4.1), which offer crucial information for replicating

the experiments and understanding the results. Next, we present the main results (Subsection 4.2), comparing the performance of our proposed method with existing MLLMs. We also conduct ablation studies to validate the effectiveness of each component of our PGPL framework (Subsection 4.3). Additionally, we examine the impact of different types of pre-trained models on the overall performance (Subsections 4.4, and 4.5). Finally, we explore the scalability of the PGPL framework (Subsection 4.6).

4.1 Implementation details

This subsection offers a detailed description of the datasets used in the experiments, the evaluation metrics adopted to assess model performance, the baseline models for comparison, and the hyperparameter settings applied in our experiments.

Datasets. We conducted extensive experiments on six popular datasets to validate the effectiveness of our framework in spatial awareness tasks and more general tasks. Among these, MME [71] and MM-Vet [72] are widely used to evaluate the general abilities of MLLMs. The MME dataset encompasses ten perceptual tasks: existence, count, position, color, poster, celebrity, scene, landmark, artwork, and OCR, with the position task specifically designed to assess the model's abilities in spatial awareness. This dataset includes 957 images and 1914 question-answer pairs, with each image associated with two corresponding questions and answers. The MM-Vet dataset includes six core tasks in computer vision and natural language processing: recognition, knowledge, OCR, spatial awareness, language generation, and mathematics. Within MM-Vet, the spatial awareness tasks are particularly aimed at evaluating the model's spatial awareness abilities. This dataset contains 200 images and 218 questions. The RefCOCO [73], RefCOCO+ [73], and RefCOCOg [74] datasets are utilized to evaluate our model's performance in referring expression comprehension. Both RefCOCO and RefCOCO+ were created through a two-player game. RefCOCO+ is specifically tailored to exclude spatial relations. In contrast, RefCOCOg includes spatial relations and features longer expressions on average. Our model's performance in image captioning is evaluated on the Flickr30k [75] dataset, which consists of 31000 images. Each image is accompanied by five textual captions provided by human annotators. These descriptions detail the scenes, objects, and activities depicted in the images.

Evaluation metrics. This paper evaluates the performance of our method using the evaluation metrics proposed in each dataset. Specifically, for the MME dataset, the model's output is limited to two types (yes or no), making it convenient to measure accuracy (based on each question) and accuracy+ (based on each image where both of the two questions need to be answered correctly) metrics. We choose to use the sum of accuracy and accuracy+ to calculate the task score. In the case of the MM-Vet dataset, GPT-4 provides specific scores to evaluate the model's performance based on existing scoring instances and the model's output under the input question and real answer conditions for each sample. The evaluation metric for RefCOCO, RefCOCO+, and RefCOCOg is accuracy, while for Flicker30K, it is CIDEr [76].

Baselines. We conducted extensive experiments to validate the effectiveness of our proposed method. The selected baseline models include LLaMA-AdapterV2-7B [77], InstructBLIP-14B [78], MiniGPT-4-14B [23], Otter-9B [79], GPT-4 Vision [80], Ferret [81], and GPT-4o [82]. Further, a series of experiments were conducted focusing on five baseline models—BLIP-2-12B [3], LLaVA-13B [2] (LLaVA-1.5), Kosmos-2 [32], GPT4RoI [31], and Shikra [36]—to specifically assess the PGPL framework's efficacy. These models were chosen for their widespread use in prior research and their status as representative benchmarks [83–85], allowing us to test our method against these well-regarded models and thus demonstrate the effectiveness and universality of our approach.

Hyperparameter settings. The hyperparameters used in this paper are identical to those of each baseline model. The key distinction lies in integrating the geometric absolute and relative position information from spatial location and scene graph generation models into the PGPL framework, following the prompt format designed to execute relevant tasks. The spatial position generation model employs the classic and user-friendly Faster R-CNN 101 [67]. The scene graph generation model is implemented using PSG [68]. The threshold of cosine similarity in the entity matching model was set to 0.8.

4.2 Main results

We systematically evaluated the performance of our proposed PGPL framework in enhancing spatial awareness abilities on specialized tasks of the MME and MM-Vet datasets. The experimental results for general (LLaMA-AdapterV2-7B, InstructBLIP-14B, BLIP-2-7B, BLIP-2-12B, LLaVA-7B, LLaVA-13B, GPT-4 Vision, Ferret, GPT-4o) and spatially aware (Kosmos-2, GPT4RoI, and Shikra) MLLMs are summarized in Table 1 [2,3,23,31,32,36,77–82]. The results show that PGPL consistently outperforms existing methods, with significant improvements across all baseline models.

Table 1 Experimental results on the MME (position task) and MM-Vet (spatial awareness task) datasets. Blod highlights the performance of our method compared to various baseline MLLMs. ↑ denotes the relative improvement percentage over the baseline.

Model	Position	Spatial
MiniGPT-4-14B [23]	81.67	22.2
Otter-9B [79]	86.67	19.3
LLaMA-AdapterV2-7B [77]	56.67	16.6
Ours (LLaMA-AdapterV2-7B)	**60.00 (↑5.9%)**	**17.8 (↑7.2%)**
InstructBLIP-14B [78]	66.67	21.1
Ours (InstructBLIP-14B)	**73.33 (↑10.0%)**	**22.9 (↑8.5%)**
BLIP-2-7B [3]	55.00	14.8
Ours (BLIP-2-7B)	**61.67 (↑12.1%)**	**16.5 (↑11.5%)**
BLIP-2-12B [3]	73.33	16.2
Ours (BLIP-2-12B)	**87.54 (↑19.4%)**	**20.1 (↑24.1%)**
LLaVA-7B [2]	53.33	24.3
Ours (LLaVA-7B)	**81.67 (↑53.1%)**	**27.1 (↑11.5%)**
LLaVA-13B [2]	133.33	24.3
Ours (LLaVA-13B)	**153.33 (↑15.0%)**	**26.5 (↑8.6%)**
GPT-4 Vision [80]	95.00	12.4
Ours (GPT-4 Vision)	**133.33 (↑40.3%)**	**14.8 (↑19.4%)**
Ferret [81]	96.67	13.8
Ours (Ferret)	**128.33 (↑32.8%)**	**15.1 (↑9.4%)**
GPT-4o [82]	180.00	58.2
Ours (GPT-4o)	**183.33(↑1.8%)**	**59.1 (↑1.5%)**
Kosmos-2 [32]	81.67	21.65
Ours (Kosmos-2)	**84.33 (↑3.3%)**	**22.42 (↑3.6%)**
GPT4RoI [31]	84.33	22.15
Ours (GPT4RoI)	**86.67 (↑2.8%)**	**23.23 (↑4.9%)**
Shikra [36]	81.67	21.40
Ours (Shikra)	**84.33 (↑3.3%)**	**22.65 (↑5.8%)**

PGPL achieved an average improvement of 16.5% across diverse models on the MME dataset. Significant gains were observed in models such as LLaVA-7B (↑53.1%), GPT-4 Vision (↑40.3%), and Ferret (↑32.8%), with notable improvements also seen in BLIP-2-12B (↑19.4%), LLaVA-13B (↑15.0%), BLIP-2-7B (↑12.1%), InstructBLIP-14B (↑10.0%), LLaMA-AdapterV2-7B (↑5.9%), and GPT-4o (↑1.8%). Even models with intrinsic spatial awareness capabilities, including Kosmos-2 (↑3.3%), Shikra (↑3.3%), and GPT4RoI (↑2.8%), benefited from PGPL. These results highlight PGPL's ability to address fundamental limitations in general multimodal models while further enhancing models already specialized in spatial reasoning.

On MM-Vet, PGPL delivered consistent improvements across all tested architectures. Performance gains were particularly strong for smaller models like BLIP-2-7B (↑11.5%), LLaVA-7B (↑11.5%), and LLaMA-AdapterV2-7B (↑7.2%), while larger models such as BLIP-2-12B (↑24.1%), GPT-4 Vision (↑19.4%), Ferret (↑9.4%), LLaVA-13B (↑8.6%), InstructBLIP-14B (↑8.5%), and GPT-4o (↑1.5%) also showed substantial improvements. Spatially aware models, including Shikra (↑5.8%), GPT4RoI (↑4.9%), and Kosmos-2 (↑3.6%), demonstrated enhanced spatial reasoning capabilities. This underscores PGPL's generalizability across model scales and architectures, with particularly pronounced effects on less optimized baselines.

The analysis highlights PGPL's versatility: it significantly improves spatial awareness in general MLLMs by incorporating GAP and GRP information without requiring additional model retraining. On the other hand, its limited but consistent impact on spatially-aware models suggests that PGPL can complement existing spatial reasoning capabilities, further validating its universal applicability. The pronounced gains in weaker models, such as LLaVA-7B and BLIP-2-7B, underscore PGPL's ability to address intrinsic deficiencies in baseline architectures, making it a valuable enhancement across a spectrum of tasks.

To further validate the practical effectiveness of the PGPL framework, we present an in-depth analysis of representative case studies involving spatial reasoning tasks. These cases demonstrate how the integration of GAP and GRP enables MLLMs to overcome various challenges in spatial understanding. Specifically, GAP provides precise geometric anchors, while GRP captures relative positional relationships. Their seamless fusion within the large model empowers it to infer higher-order spatial and semantic reasoning, addressing limitations inherent in baseline MLLMs.

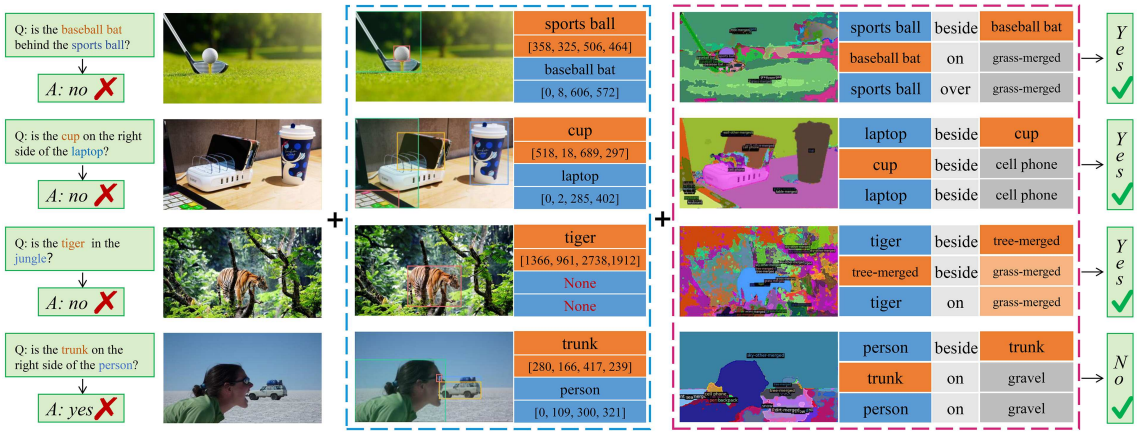

Figure 4 (Color online) Illustration of representative cases analyzed in the study. From top to bottom, the figure sequentially represents Cases 1 to 4. The magenta-colored bounding boxes highlight GRP relationships, while the blue-colored bounding boxes represent GAP information. This visualization demonstrates how GAP and GRP work in tandem to resolve spatial reasoning challenges.

As shown in Figure 4, in Case 1, the question concerned whether the baseball bat was behind the sports ball. The baseline model failed due to incomplete geometric information and a lack of relational understanding. GRP extracted relative relationships such as "beside" and "on" between key objects in the scene, while GAP, despite missing some bounding boxes, anchored the position of the sports ball. By leveraging these complementary inputs, the large model accurately deduced that the baseball bat was indeed behind the sports ball, demonstrating the framework's ability to combine relational reasoning with geometric precision.

Case 2 highlighted PGPL's ability to resolve ambiguities caused by closely positioned objects, such as a cup and a laptop. The baseline model struggled to differentiate between nearby objects due to overlapping features. GRP provided contextual relational cues, identifying the spatial hierarchy (e.g., "beside" relationships among the cup, laptop, and other nearby objects). GAP reinforced this by supplying bounding box coordinates, enabling the large model to determine that the cup was correctly located to the right of the laptop.

Case 3 showcased the framework's robustness in handling noisy or generalized entity labels, such as tree-merged and grass-merged. GRP effectively captured the relative spatial relationships, such as "beside" and "on", even when absolute position data was incomplete or ambiguous. GAP, though limited in providing precise bounding boxes, still offered geometric anchors for reasoning. By synthesizing these two sources of information, the large model was able to infer the intended spatial relationships, underscoring the framework's resilience in scenarios with imprecise or noisy entity data.

Finally, Case 4 illustrated PGPL's capability to mitigate optical illusions and scale-based misinterpretations. In a scenario where a person appeared larger than a vehicle, GRP provided relational context (e.g., "on gravel"), while GAP revealed that the bounding box sizes were inconsistent with the perceived scale. This discrepancy allowed the large model to correctly infer that the trunk was not on the right side of the person. This case highlights how GAP's geometric data serves as a critical anchor for resolving visual ambiguities, while GRP contextualizes the scene's relational structure.

Together, these case studies demonstrate the complementary strengths of GAP and GRP. GAP provides precise geometric information critical for accurate absolute positioning, while GRP captures essential relational dynamics between entities. When fused within the PGPL framework, these inputs enable the large model to achieve high-order spatial reasoning, even in complex or deceptive visual scenarios. The ability to handle ambiguous contexts, incomplete data, and challenging visual environments makes the PGPL framework a versatile and robust enhancement for MLLMs. This capability is particularly significant for applications requiring advanced spatial understanding, such as autonomous driving, robotic navigation, and multimodal question answering. The results reaffirm the effectiveness of PGPL in elevating spatial reasoning across diverse and challenging tasks.

4.3 Ablation studies

4.3.1 *Performance analysis*

We conducted a comprehensive set of ablation experiments using two different baseline models, BLIP-2-12B and LLaVA-13B, to validate the effectiveness of GAP, GRP, and their combined application (GAP + GRP, PGPL framework). The prompt designs and the corresponding ablation experimental results for these tests are detailed

Table 2 Design details of prompt for multidimensional information fusion in ablation analysis.

Model	Prompt design
Baseline model + GAP	The targets along with their geometric absolute positions in the image are as follows: $\text{Entity}_1 : (x_1^*, y_1^*, x_1^\dagger, y_1^\dagger), \ldots, \text{Entity}_k : (x_k^*, y_k^*, x_k^\dagger, y_k^\dagger)$. Please answer the following questions based on the above information and the image itself: QUESTION, and directly tell me the answer that you think is correct.
Baseline model + GRP	The scene in the image has the following relationships $(e_{h1}, r_1, e_{t1}), \ldots, (e_{hz}, r_z, e_{tz})$. Please answer the following questions based on the above information and the image itself: QUESTION, and directly tell me the answer that you think is correct.
Baseline model + GAP + GRP (PGPL)	The scene in the image has the following relationships $(e_{h1}, r_1, e_{t1}), \ldots, (e_{hz}, r_z, e_{tz})$, and the targets along with their geometric positions are as follows: $\text{Entity}_1 : (x_1^*, y_1^*, x_1^\dagger, y_1^\dagger), \ldots, \text{Entity}_k : (x_k^*, y_k^*, x_k^\dagger, y_k^\dagger)$. Please answer the following questions based on the above information and the image itself: QUESTION, and directly tell me the answer that you think is correct.

Table 3 Ablation analysis on the MME and MM-Vet datasets, using BLIP-2-12B and LLaVA-13B as baseline models. Blod indicates the best performance, while ↑ denotes the percentage of performance improvement over the baselines.

Model	Position	Spatial
BLIP-2-12B [3]	73.33	16.2
BLIP-2-12B + GAP	78.36 (↑6.9%)	18.8 (↑16.0%)
BLIP-2-12B + GRP	80.48 (↑9.6%)	19.6 (↑21.0%)
BLIP-2-12B + GAP + GRP	**87.54 (↑19.4%)**	**20.1 (↑24.1%)**
LLaVA-13B [2]	133.33	24.3
LLaVA-13B + GAP	143.33 (↑7.5%)	24.8 (↑2.1%)
LLaVA-13B + GRP	143.33 (↑7.5%)	25.4 (↑4.5%)
LLaVA-13B + GAP + GRP	**153.33 (↑15.0%)**	**26.5 (↑8.6%)**

in Tables 2 and 3, and Figure 5, respectively.

Focusing on the MME dataset's position task, integrating GAP with BLIP-2-12B led to a 6.9% improvement in the score, from 73.33 to 78.36. In comparison, LLaVA-13B showed a 7.5% increase, from 133.33 to 143.33. The GAP mechanism effectively provides precise location anchors for each object, mitigating the ambiguity in absolute object positioning and significantly enhancing the model's ability to discern individual spatial attributes. Further, the incorporation of GRP with BLIP-2-12B and LLaVA-13B achieved a 9.6% and 7.5% score increase for the position task, respectively. GRP excels in modeling intricate spatial relationships among objects, particularly in complex multi-object scenarios. It enhances the model's ability to infer relative spatial dynamics, complementing the GAP mechanism. The synergistic effect of combining both GAP and GRP with BLIP-2-12B resulted in a significant 19.4% improvement, from 73.33 to 87.54. A similar fusion approach with LLaVA-13B yielded a 15.0% enhancement, from 133.33 to 153.33. These results highlight the complementary nature of GAP and GRP: GAP provides absolute positional anchors, while GRP focuses on the relational geometry, together enabling a multi-scale spatial understanding. Trends observed on the MM-Vet dataset were paralleled on the MME dataset, with varying degrees of improvement, which affirms the adaptability and effectiveness of our PGPL framework across different spatial awareness tasks.

We further substantiate the efficacy of the GAP and GRP information proposed in our method through qualitative experimental results, as depicted in Figure 5. In scenario (1), the existing BLIP-2-12B model fails to accurately discern the spatial relationship between the white mouse and the black keyboard, incorrectly predicting that the white mouse is to the left of the black keyboard. However, by employing a spatial position generation model, we accurately capture the detailed categories and geometric absolute positions of all mice and keyboards in the image, enabling the model to correctly determine that the white mouse is on the right side of the black keyboard. In scenario (2), the existing BLIP-2-12B model cannot correctly identify the spatial relationship between the sheep and the tree, suggesting that the sheep is not in front of the tree. In contrast, our scene graph generation model accurately captures the geometric relative position information related to the sheep and trees in the image, thereby assisting the model in making the correct inference that the sheep is indeed in front of the tree.

4.3.2 *Inference time analysis*

To address the potential computational overhead introduced by PGPL, we conducted extensive experiments across six representative models, including MiniGPT-4-14B, BLIP-2-12B, LLaVA-13B, Kosmos-2, GPT4RoI, and Shikra. Our analysis reveals that while PGPL avoids incremental training, the additional modules (GAP and GRP) introduce a modest increase in inference time. Specifically, the average total response time increases by 17.2% in the default serial mode. However, by implementing parallel preprocessing techniques (where TEE, GAP, and GRP operations are executed concurrently), this overhead is significantly reduced to 8.3%, demonstrating the efficiency of our approach.

To provide a detailed breakdown, we analyze GPT4RoI as a representative case. The baseline model exhibits a

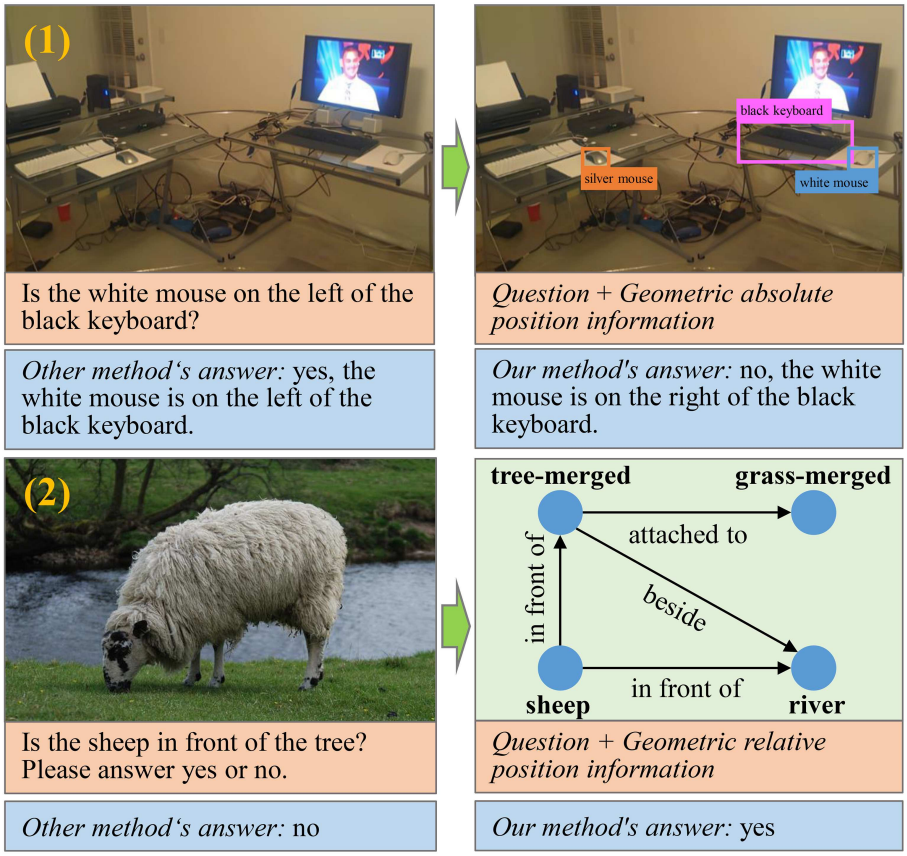

Figure 5 (Color online) Visual results from ablation studies highlighting the effectiveness of our model in discerning spatial relationships.

response time of 4289 ms, which increases to 5265 ms when PGPL is integrated in serial mode. The time distribution across PGPL stages is as follows:

- TEE: 4% (210 ms);
- GAP: 6% (316 ms);
- EM_1 (first entity matching): 1% (53 ms);
- GRP: 7% (369 ms);
- EM_2 (second entity matching): 1% (53 ms);
- Model inference: 81% (4264 ms).

Notably, the majority of the overhead is attributed to GAP and GRP (13% combined), while entity matching (EM_1 and EM_2) contributes minimally (2%), indicating that the PGPL pipeline is efficiently implemented. This 8.3% average time increase is a reasonable trade-off given the significant performance improvement, making PGPL highly practical for real-world applications.

Furthermore, we believe that future optimizations could significantly enhance the module's efficiency, reduce response times, and improve system real-time performance and user experience. For instance, replacing GAP and GRP with lightweight expert models could reduce their time contribution by 30%–50%, while leveraging GPU-optimized libraries for parallel preprocessing could further reduce computational overhead.

4.4 The impact and selection of pre-trained models in the PGPL framework

Our experimental results demonstrate that the SPG and SGG models in the PGPL framework significantly improve the spatial awareness capabilities of MLLMs. Building on this evidence, we investigate two critical questions regarding pre-trained models. (1) Does a more powerful pre-trained model correlate with better overall performance? (2) Is the performance increase worth compared to the number of parameters introduced by the pre-trained model?

The experimental results examining the relationship between the performance of pre-trained models and overall effectiveness are presented in Table 4 [3,60,68,86–88]. For spatial position generation models (SPG, rows 3–6), there is a clear trend observed: from Faster R-CNN 50 to Co-DETR, as the pre-trained model's strength increases, there is a corresponding enhancement in overall performance for both position and spatial awareness tasks. Similarly, for

Table 4 Impact of SPG and SGG models with varying performance on the PGPL framework.

Model	Position	Spatial
BLIP-2-12B [3]	73.33	16.2
BLIP-2-12B + Faster R-CNN 50	76.67	17.5
BLIP-2-12B + Faster R-CNN 101	78.36	18.8
BLIP-2-12B + Faster R-CNN 152	81.67	19.7
BLIP-2-12B + Co-DETR [60]	81.67	20.8
BLIP-2-12B + VCTree [86]	76.67	17.8
BLIP-2-12B + MOTIFS [87]	76.67	18.3
BLIP-2-12B + GPSNet [88]	76.67	19.1
BLIP-2-12B + PSG [68]	81.67	19.6

Table 5 Comparative results of different types of spatial position generation models (object detection with Faster R-CNN, instance segmentation with ISBNet, FastInst, etc.).

Model	Position	Spatial
BLIP-2-12B [3]	73.33	16.2
BLIP-2-12B + Faster R-CNN 101	78.36	18.8
BLIP-2-12B + Faster R-CNN 152	81.67	19.7
BLIP-2-12B + ISBNet [93]	75.00	19.1
BLIP-2-12B + FastInst [94]	78.36	18.6
BLIP-2-12B + YOSO [95]	81.67	18.1
BLIP-2-12B + ODISE [62]	81.67	18.4

scene graph generation models (SGG, rows 7–10), ranging from VCTree to PSG, an augmentation in the pre-trained model's abilities leads to continuous improvements in overall performance. Hence, the more powerful the pre-trained model, the more accurate the spatial awareness outcomes and the greater the increase in performance.

We investigate the correlation between the number of parameters and performance improvements in the PGPL framework, focusing on the SPG model Co-DETR and the SGG model PSG. Co-DETR adds 348 M parameters, a 2.9% increase from the BLIP-2-12B baseline, and improves performance by 11.4% in position and 28.4% in spatial awareness tasks. PSG, introducing 603 M parameters (5.0% increase), enhances performance by 11.4% in position and 21.0% in spatial awareness tasks. These results indicate that performance benefits significantly outweigh the parameter count increase. Furthermore, the adoption of lightweight pre-trained models [89, 90] can reduce the number of parameters while preserving the accuracy of geometric spatial information. Crucially, our method enhances spatial awareness without the need for retraining the MLLM, thereby effectively utilizing existing models to address spatial challenges.

4.5 Space position generation model analysis

We have established selection criteria for space position generation models and verified their impact on overall performance. Based on overall needs, the selected model must satisfy two essential criteria; (1) the model must accurately capture the type information of entities within images; (2) the model must precisely generate the detailed position information of the corresponding entities. Taking image segmentation models as an example, semantic segmentation models cannot meet these requirements because they generate a label for each pixel, assigning the same label to entities of the same type, thus being unable to distinguish between different entities. However, instance segmentation [91] and panoptic segmentation [92] models can meet these requirements.

We validated the performance of different spatial position generation models, including object detection and instance segmentation. As detailed in Table 5 [3, 62, 93–95], we compared the results of using instance segmentation models, which capture a quadrilateral of an object from 4 points, against object detection models that also obtain quadrilateral spatial positions. The results indicate that both model types can enhance MLLM performance in spatial awareness. Consequently, researchers are encouraged to explore additional spatial position generation techniques, such as panoptic segmentation and depth estimation algorithms, to further their research objectives.

We further investigate instance segmentation models. Our objective was to ascertain the optimal number of segmentation points that would facilitate the most accurate segmentation without superfluous computational overhead. As depicted in Figure 6, we conducted a thorough assessment of several leading instance segmentation models (ISBNet [93], FastInst [94], YOSO [95], and ODISE [62]) across the position and spatial awareness tasks in the MME and MM-Vet datasets, respectively. A methodical examination of these models, utilizing varying segmentation points

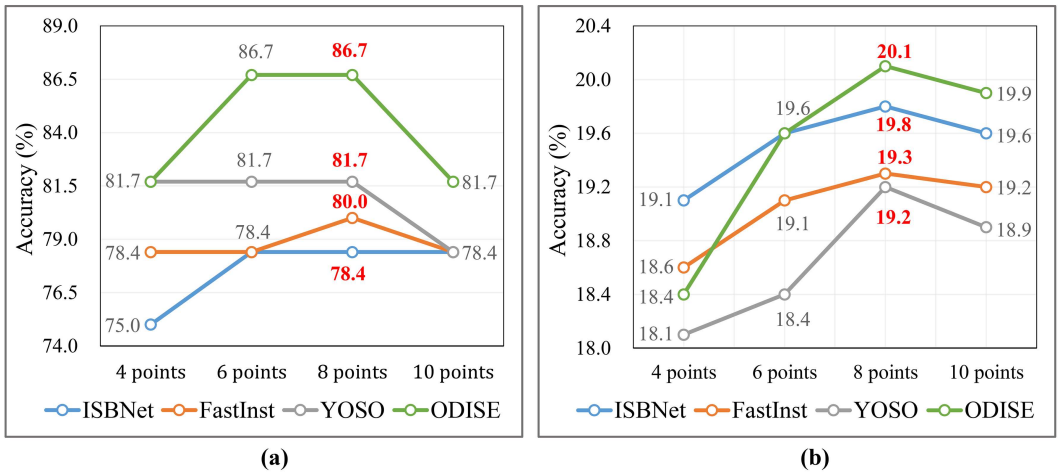

Figure 6 (Color online) Experimental results for selecting optimal segmentation points in instance segmentation models. (a) MME position task; (b) MM-Vet spatial awareness task.

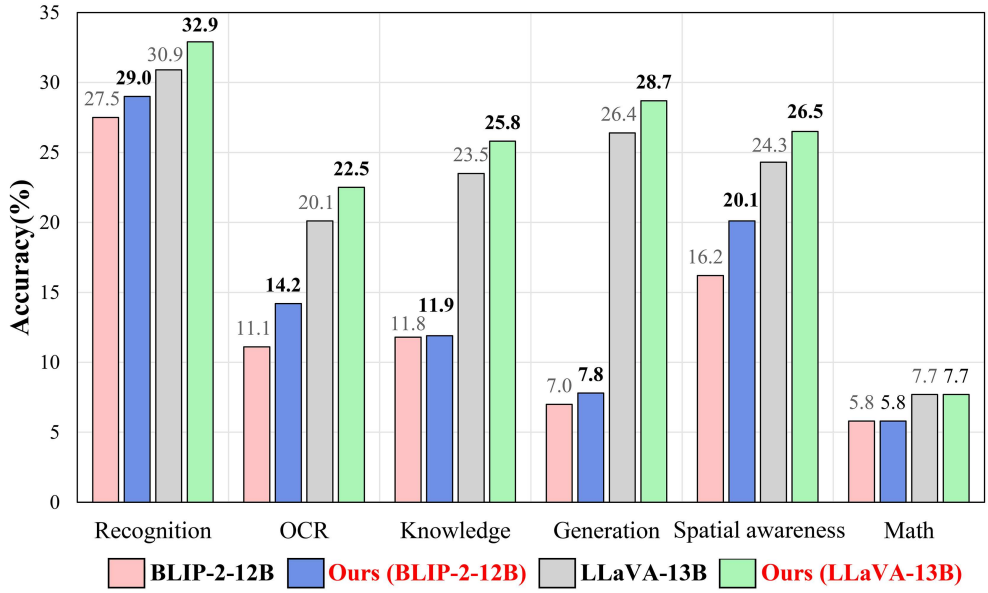

Figure 7 (Color online) Experimental results on six core vision-language tasks in the MM-Vet dataset.

(4, 6, 8, and 10), revealed a consistent trend: optimal performance was achieved with 8 segmentation points. This finding underscores a point of diminishing returns where additional segmentation points cease to yield proportional performance improvements. Thus, the 8-point setup is identified as the most efficient, offering a balance between accuracy and computational efficiency.

4.6 Scalability of the PGPL framework

In addition to evaluating the PGPL framework's performance on two tasks specifically designed to assess spatial awareness, we also tested its effectiveness across a significantly broader range of tasks. This evaluation includes six core visual-language tasks within the MM-Vet dataset (Figure 7), ten perceptual tasks in the MME dataset (Table 6), as well as traditional tasks such as referring expression comprehension and image captioning (Table 7).

The experiment results on the MM-Vet dataset are illustrated in Figure 7. It is evident from the figure that our method, using BLIP-2-12B and LLaVA-13B as baseline models, achieves significant improvements in the first five core visual-language tasks. For instance, in the recognition and OCR tasks, our method shows an increase of 6.9% and 11.9%, respectively, over LLaVA-13B. The lack of improvement in the math task can be attributed to our method not including enhanced abilities for mathematical computation.

The experiment results on the MME dataset are shown in Table 6. The results indicate significant performance

Table 6 Performance on ten tasks in the MME dataset. 'exist' denotes existence, 'celeb' for celebrity, and 'art' for artwork. Blod indicates our method's performance surpasses the baseline.

Model	Exist	Count	Position	Color	Poster	Celeb	Scene	Landmark	Art	OCR
BLIP-2-12B [3]	160.00	135.00	73.33	145.00	141.84	105.59	145.25	137.25	136.50	110.00
Ours (BLIP-2-12B)	**168.00**	**140.00**	**87.54**	**148.33**	141.84	105.59	**147.98**	**138.00**	136.50	**125.00**
LLaVA-13B [2]	185.00	155.00	133.33	170.00	160.54	152.94	161.25	170.50	117.75	125.00
Ours (LLaVA-13B)	**190.00**	**160.00**	**153.33**	**180.00**	160.54	152.94	**164.50**	**172.25**	117.75	**140.00**

Table 7 Performance of our PGPL framework on referring expression comprehension and image captioning tasks. The bolded values represent the results after using the PGPL framework.

Model		RefCOCO			RefCOCO+			RefCOCOg		Flickr30k
		Val	TestA	TestB	Val	TestA	TestB	Val	TestA	
General MLLMs	BLIP-2-12B [3]	55.8	60.9	49.7	47.6	52.8	44.9	62.3	63.4	79.1
	Ours (BLIP-2-12B)	**58.1**	**62.9**	**51.4**	**49.8**	**54.6**	**46.4**	**64.6**	**65.8**	**80.7**
	LLaVA-13B [2]	88.1	92.3	82.0	83.5	88.4	75.2	83.0	84.1	74.5
	Ours (LLaVA-13B)	**89.5**	**93.6**	**83.8**	**85.1**	**89.9**	**76.9**	**84.7**	**85.5**	**76.0**
Specific MLLMs	Kosmos-2 [32]	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7	77.8
	Ours (Kosmos-2)	**54.7**	**59.8**	**49.3**	**47.5**	**52.9**	**44.2**	**62.8**	**63.9**	**79.2**
	GPT4RoI [31]	54.1	59.3	48.8	46.3	51.7	43.5	61.2	62.3	78.4
	Ours (GPT4RoI)	**56.4**	**61.5**	**50.6**	**48.4**	**53.8**	**45.6**	**63.5**	**64.7**	**80.1**
	Shikra-13B [36]	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2	73.9
	Ours (Shikra-13B)	**89.2**	**92.4**	**83.7**	**84.5**	**89.1**	**76.2**	**84.0**	**84.5**	**75.3**

enhancements in tasks (existence, count, position, color, scene, landmark, and OCR) related to recognition and scene understanding. For instance, in the color and OCR tasks, our method's performance achieved improvements of 5.9% and 12.0%, respectively, over LLaVA-13B. Notably, tasks such as poster, celebrity, and artwork did not see performance gains, due to our method not incorporating recognition abilities specific to these areas. However, even without improvements, our method does not degrade the performance of these tasks, as it avoids introducing noise by not attempting to process unrelated entities.

To evaluate the extensive applicability of our method, we also assessed our model's performance on more traditional tasks, with results presented in Table 7. The data reveal that our approach yields significant enhancements across three referring expression comprehension datasets (RefCOCO, RefCOCO+, RefCOCOg) and one image captioning dataset (Flickr30k). This improvement is consistent whether applied to general MLLMs (BLIP-2-12B, LLaVA-13B) or MLLMs enhanced with specific spatial awareness abilities (Kosmos-2, GPT4RoI, Shikra-13B). In summary, our proposed PGPL framework not only enhances the spatial awareness capabilities of MLLM but also significantly improves its performance across a broader range of tasks. Consequently, the PGPL framework establishes itself as a potent instrument for expanding the versatility and augmenting the overall efficacy of MLLM.

5 Conclusion

We present the PGPL (precise geometric position learning) framework, a novel approach designed to enhance the spatial awareness capabilities of MLLMs. The PGPL framework leverages pre-trained spatial position generation models and scene graph generation models to provide both absolute and relative geometric position information between target entities. By integrating multidimensional spatial information, PGPL enables MLLMs to more accurately address user queries related to spatial awareness. To validate the effectiveness of our framework, we conducted extensive experiments on six benchmark datasets, comparing PGPL against five state-of-the-art baselines. The experimental results demonstrate that PGPL significantly improves the performance of MLLMs in spatial awareness tasks, underscoring its potential as a robust solution for enhancing spatial reasoning in multimodal systems. Limitations: when existing pre-trained models are unable to accurately identify target entities in specific domains, the effectiveness of our framework may be impacted. To address this challenge, we plan to explore universal strategies for target entity recognition in specific domains in future research. One preliminary strategy we are considering involves retraining smaller models with a limited amount of domain-specific data. Broader impact: this study highlights the potential of integrating pre-trained smaller models with large-scale MLLMs to achieve task-specific performance enhancements in a cost-effective manner. By demonstrating the efficacy of this hybrid approach, we aim to inspire further research into leveraging smaller, domain-specific models to augment

the capabilities of large multimodal systems. We believe that our work opens new avenues for improving spatial awareness and related tasks in MLLMs, with potential applications in fields such as robotics, autonomous navigation, and human-computer interaction.

Acknowledgements This work was supported by National Science and Technology Major Project (Grant No. 2020AAA0109401) and National Natural Science Foundation of China (Grant No. 62192731). We would like to express our gratitude to Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, for providing computing resources.

References

1 Wang X, Chen G, Qian G, et al. Large-scale multi-modal pre-trained models: a comprehensive survey. Mach Intell Res, 2023, 20: 447–482
2 Liu H, Li C, Wu Q, et al. Visual instruction tuning. 2023. ArXiv:2304.08485, 2023
3 Li J, Li D, Savarese S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. 2023. ArXiv:2301.12597
4 Zhang Z, Zhang A, Li M, et al. Multimodal chain-of-thought reasoning in language models. 2023. ArXiv:2302.00923
5 Wang W, Chen Z, Chen X, et al. VisionLLM: large language model is also an open-ended decoder for vision-centric tasks. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
6 Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ, 2023, 103: 102274
7 Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiol, 2023, 1: 100017
8 Taori R, Gulrajani I, Zhang T, et al. Alpaca: a strong, replicable instruction-following model. 2023. https://crfm.stanford.edu/2023/03/13/alpaca.html
9 Chiang W L, Li Z, Lin Z, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90% ChatGPT quality. 2023. https://lmsys.org/blog/2023-03-30-vicuna/
10 Bae S, Kyung D, Ryu J, et al. EHRXQA: a multi-modal question answering dataset for electronic health records with chest X-ray images. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
11 Lin W, Chen J, Mei J, et al. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
12 Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. 2023. ArXiv:2306.13549
13 Chang Y, Ko Y. Two-step masked language model for domain-adapting multi-modal task-oriented dialogue systems. IEEE ACM Trans Audio Speech Lang Process, 2024, 32: 2938–2943
14 Majumdar A, Yadav K, Arnaud S, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? In: Proceedings of the Advances in Neural Information Processing Systems, 2024
15 Wang J, Wu Z, Li Y, et al. Large language models for robotics: opportunities, challenges, and perspectives. 2024. ArXiv:2401.04334
16 Mañas O, Krojer B, Agrawal A. Improving automatic VQA evaluation using large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 4171–4179
17 Khan Z, BG V K, Schulter S, et al. Exploring question decomposition for zero-shot VQA. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
18 Gao J, Wu Q, Blair A, et al. LoRA: a logical reasoning augmented dataset for visual question answering. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
19 Qian T, Chen J, Zhuo L, et al. Nuscenes-QA: a multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 4542–4550
20 Tian X, Jiang T, Yun L, et al. OCC3D: a large-scale 3D occupancy prediction benchmark for autonomous driving. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
21 Chen Q, Ye H, Hong Y. Med3dinsight: enhancing 3D medical image understanding with 2D multi-modal large language models. 2024. ArXiv:2403.05141
22 Muhammad G, Alshehri F, Karray F, et al. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Inf Fusion, 2021, 76: 355–375
23 Zhu D, Chen J, Shen X, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models. 2023. ArXiv:2304.10592
24 Kamath A, Hessel J, Chang K W. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023
25 Yin Z, Wang J, Cao J, et al. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
26 Zhang A, Fei H, Yao Y, et al. VPGTrans: transfer visual prompt generator across LLMs. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
27 Dong Q, Li L, Dai D, et al. A survey for in-context learning. 2022. ArXiv:2301.00234
28 Lu P, Peng B, Cheng H, et al. Chameleon: plug-and-play compositional reasoning with large language models. 2023. ArXiv:2304.09842
29 Wang Q C, Xiao Z, Mao Y, et al. Model predictive task sampling for efficient and robust adaptation. 2025. ArXiv:2501.11039
30 Ye Q, Xu H, Xu G, et al. mPLUG-Owl: modularization empowers large language models with multimodality. 2023. ArXiv:2304.14178
31 Zhang S, Sun P, Chen S, et al. GPT4ROI: instruction tuning large language model on region-of-interest. 2023. ArXiv:2307.03601
32 Peng Z, Wang W, Dong L, et al. Kosmos-2: grounding multimodal large language models to the world. 2023. ArXiv:2306.14824
33 Shen Y, Song K, Tan X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in huggingface. 2023. ArXiv:2303.17580
34 Zhao Y, Li Z, Zhang F, et al. Enhancing subtask performance of multi-modal large language model. 2023. ArXiv:2308.16474
35 Wu C, Yin S, Qi W, et al. Visual ChatGPT: talking, drawing and editing with visual foundation models. 2023. ArXiv:2303.04671
36 Chen K, Zhang Z, Zeng W, et al. Shikra: unleashing multimodal LLM's referential dialogue magic. 2023. ArXiv:2306.15195
37 Zhang W, Lin T, Liu J, et al. HyperLLaVA: dynamic visual and language expert tuning for multimodal large language models. 2024. ArXiv:2403.13447
38 Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: a survey. Proc IEEE, 2023, 111: 257–276
39 Pu Y, Liang W, Hao Y, et al. Rank-DETR for high quality object detection. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
40 Meng L, Dai X, Yang J, et al. Learning from rich semantics and coarse locations for long-tailed object detection. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
41 Yu Y, Wang C, Fu Q, et al. Techniques and challenges of image segmentation: a review. Electronics, 2023, 12: 1199
42 Wang X, Li S, Kallidromitis K, et al. Hierarchical open-vocabulary universal image segmentation. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
43 Wang H, Li X. Towards generic semi-supervised framework for volumetric medical image segmentation. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
44 Monnier Q, Pouli T, Kpalma K. Survey on fast dense video segmentation techniques. Comput Vision Image Understand, 2024, 241: 103959

45 Wu F, Marquez-Neila P, Zheng M, et al. Correlation-aware active learning for surgery video segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024. 2010–2020
46 Kong L, Xie S, Hu H, et al. Robodepth: robust out-of-distribution depth estimation under corruptions. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
47 Saxena S, Herrmann C, Hur J, et al. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
48 Li R, Zhang S, Wan B, et al. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 11109–11119
49 Li H, Zhu G, Zhang L, et al. Scene graph generation: a comprehensive survey. Neurocomputing, 2024, 566: 127052
50 Zhao M, Zhang L, Wang W, et al. Adversarial attacks on scene graph generation. IEEE Trans Inform Forensic Secur, 2024, 19: 3210–3225
51 Lu J, Chen L, Guan H, et al. Improving rare relation inferring for scene graph generation using bipartite graph network. Comput Vision Image Understand, 2024, 239: 103901
52 Zhang L, Zhai X, Zhao Z, et al. What if the TV was off? Examining counterfactual reasoning abilities of multi-modal language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 4629–4633
53 Li L, Lei J, Gan Z, et al. Adversarial VQA: a new benchmark for evaluating the robustness of VQA models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 2042–2051
54 Yang Z, Gan Z, Wang J, et al. An empirical study of GPT-3 for few-shot knowledge-based VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022. 3081–3089
55 Zhao P, Zheng S, Zhao W, et al. Rethinking two-stage referring expression comprehension: a novel grounding and segmentation method modulated by point. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 7487–7495
56 Nguyen T, Gadre S Y, Ilharco G, et al. Improving multimodal datasets with image captioning. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
57 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186
58 Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition. In: Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), 2009. 147–155
59 Li Y, Mao H, Girshick R, et al. Exploring plain vision transformer backbones for object detection. In: Proceedings of European Conference on Computer Vision, 2022. 280–296
60 Zong Z, Song G, Liu Y. Detrs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 6748–6758
61 Wu Q, Castleman K R. Image segmentation. In: Proceedings of Microscope Image Processing, 2023. 119–152
62 Xu J, Liu S, Vahdat A, et al. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 2955–2966
63 Fang Z, Guo X, Lin J, et al. An embedding-unleashing video polyp segmentation framework via region linking and scale alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 1744–1752
64 Weng Y, Han M, He H, et al. Mask propagation for efficient video semantic segmentation. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
65 Cui Z, Sheng H, Yang D, et al. Light field depth estimation for non-Lambertian objects via adaptive cross operator. IEEE Trans Circuits Syst Video Technol, 2024, 34: 1199–1211
66 Luo X, Huang J B, Szeliski R, et al. Consistent video depth estimation. ACM Trans Graph, 2020, 39: 71
67 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149
68 Yang J, Ang Y Z, Guo Z, et al. Panoptic scene graph generation. In: Proceedings of European Conference on Computer Vision, 2022. 178–196
69 Shao Z, Yu Z, Wang M, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 14974–14983
70 Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv, 2023, 55: 1–35
71 Fu C, Chen P, Shen Y, et al. MME: a comprehensive evaluation benchmark for multimodal large language models. 2023. ArXiv:2306.13394
72 Yu W, Yang Z, Li L, et al. MM-Vet: evaluating large multimodal models for integrated capabilities. 2023. ArXiv:2308.02490
73 Yu L, Poirson P, Yang S, et al. Modeling context in referring expressions. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 69–85
74 Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 11–20
75 Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 2641–2649
76 Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 4566–4575
77 Gao P, Han J, Zhang R, et al. LlaMA-Adapter V2: parameter-efficient visual instruction model. 2023. ArXiv:2304.15010
78 Dai W, Li J, Li D, et al. Instructblip: towards general-purpose vision-language models with instruction tuning. 2023. ArXiv:2305.06500
79 Li B, Zhang Y, Chen L, et al. Otter: a multi-modal model with in-context instruction tuning. 2023. ArXiv:2305.03726
80 Wu W, Yao H, Zhang M, et al. GPT4Vis: what can GPT-4 do for zero-shot visual recognition? 2023. ArXiv:2311.15732, 2023
81 You H, Zhang H, Gan Z, et al. Ferret: refer and ground anything anywhere at any granularity. In: Proceedings of the 12th International Conference on Learning Representations, 2023
82 Hurst A, Lerer A, Goucher A P, et al. GPT-4o system card. 2024. ArXiv:2410.21276
83 Cao Y, Li S, Liu Y, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. 2023. ArXiv:2303.04226
84 Li C, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. 2023. ArXiv:2306.00890
85 Lu Y, Li C, Liu H, et al. An empirical study of scaling instruct-tuned large multimodal models. 2023. ArXiv:2309.09958
86 Tang K, Zhang H, Wu B, et al. Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 6619–6628
87 Zellers R, Yatskar M, Thomson S, et al. Neural motifs: scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 5831–5840
88 Lin X, Ding C, Zeng J, et al. GPS-Net: graph property sensing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 3746–3753
89 Chen C, Liu M, Meng X, et al. Refinedetlite: a lightweight one-stage object detection framework for CPU-only devices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020. 700–701
90 Wei M, Zhan W. YOLO_MRC: a fast and lightweight model for real-time detection and individual counting of Tephritidae pests. Ecol Inf, 2024, 79: 102445

91 Le M Q, Nguyen T V, Le T N, et al. MaskDiff: modeling mask distribution with diffusion probabilistic model for few-shot instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 2874–2881
92 Sun S, Wang W, Howard A, et al. ReMaX: relaxing for better training on efficient panoptic segmentation. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
93 Ngo T D, Hua B S, Nguyen K. ISBNet: a 3D point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 13550–13559
94 He J, Li P, Geng Y, et al. Fastinst: a simple query-based model for real-time instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 23663–23672
95 Hu J, Huang L, Ren T, et al. You only segment once: towards real-time panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 17819–17829