360° Image Perception with MLLMs: A Comprehensive Benchmark and a Training-Free Method

Huyen. T. T. Tran[1], Van-Quang Nguyen[1], Farros Alferro[1], Kang-Jun Liu[1], and Takayuki Okatani[1,2]

[1] Department of System Information Sciences, Graduate School of Information Sciences, Tohoku University, Japan
[2] RIKEN Center for AIP
{tran,quang,farrosalferro,kjliu,okatani}@vision.is.tohoku.ac.jp

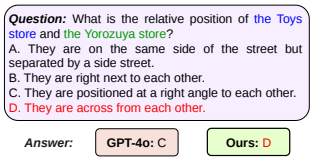
Abstract. Multimodal Large Language Models (MLLMs) have shown impressive abilities in understanding and reasoning over conventional images. However, their perception of 360° images remains largely underexplored. Unlike conventional images, 360° images capture the entire surrounding environment, enabling holistic spatial reasoning but introducing challenges such as geometric distortion and complex spatial relations. To comprehensively assess MLLMs' capabilities to perceive 360° images, we introduce **360Bench**, a Visual Question Answering (VQA) benchmark featuring 7K-resolution 360° images, seven representative (sub)tasks with annotations carefully curated by human annotators. Using 360Bench, we systematically evaluate seven MLLMs and six enhancement methods, revealing their shortcomings in 360° image perception. To address these challenges, we propose **Free360**, a training-free scene-graph-based framework for high-resolution 360° VQA. Free360 decomposes the reasoning process into modular steps, applies adaptive spherical image transformations to 360° images tailored to each step, and seamlessly integrates the resulting information into a unified graph representation for answer generation. Experiments show that Free360 consistently improves its base MLLM and provides a strong training-free solution for 360° VQA tasks. The source code and dataset will be available at https://tranhuyen1191.github.io/360Bench-Free360/.

Keywords: Multimodal Large Language Models · 360° Image Perception · Scene Graph

1 Introduction

The development of MLLMs, from proprietary to open-source models [2, 29, 35, 52], has enabled models to perform complex tasks involving the understanding and integration of multimodal information [3, 44]. In contrast to conventional models, which are typically trained on small datasets tailored to specific tasks (e.g., image captioning), MLLMs are pre-trained on massive corpora, allowing

(a) Two formats of 360° images



(b) Responses of GPT-4o and Free360

Fig. 1: (a) Example of an image from **360Bench** in two common projection formats with relevant entities highlighted. (b) Responses of GPT-4o and **Free360** (Ours) with correct answers highlighted in red. GPT-4o struggles with spatial reasoning, whereas Free360 correctly infers spatial relations. 360° image source: [15].

them to tackle a wide range of tasks via natural language instructions [14,23]. As a result, MLLMs have been increasingly applied in diverse real-world domains, such as healthcare, robotics, and education [19,30].

Unlike conventional images, 360° images offer a complete view of the surrounding environment within a single frame, enabling comprehensive spatial reasoning across the entire scene—a capability that single or even multiple conventional images cannot effectively provide. While still uncommon, enabling MLLMs to handle 360° images holds great promise for applications requiring holistic scene understanding, such as autonomous driving, assistive robotics, and surveillance.

However, realizing this potential in MLLMs comes with significant challenges. First, *spatial reasoning* over a continuous spherical space is inherently complex. As shown in Fig. 1(b), MLLMs often misinterpret spatial relations. Second, 360° images have the extensive field of view with rich visual content, requiring models to capture *fine-grained features* for accurate understanding. Third, since most models are designed for planar images, 360° images are usually projected into planar representations rather than processed in their native spherical form. Such projections cause *geometric distortion* (e.g., stretched poles from uneven sampling) and *object fragmentation* (when objects span across image boundaries). Among projection methods, *EquiRectangular Projection (ERP)* and *CubeMap Projection (CMP)* are the most commonly used [7,27], as illustrated in Fig. 1(a). ERP maps a sphere's longitude and latitude to 2D horizontal and vertical coordinates [31], whereas CMP projects the sphere onto six cube faces [6,16]. To date, it remains unclear which format facilitates MLLM perception more effectively.

Recent efforts have begun to explore question answering and reasoning over omnidirectional data [42, 48]. OmniVQA [48] introduces a 360° VQA dataset built from 1440×720 ERP images of indoor scenes and evaluates five MLLM families on two tasks derived from six templates. Very recently, concurrent work has also started to investigate broader 360° image understanding. In particular, ODI-Bench [42] studies ten VQA tasks and introduces a training-free reasoning framework for MLLMs. We view this line of work as complementary to ours. In contrast, our focus is high-resolution 360° VQA with manually curated annotations and subtasks designed to probe fine-grained perception, spatial reasoning, and robustness to projection-induced challenges.

Table 1: Comparison between 360Bench and existing VQA datasets. **CI**, **360I**, **360V**, and **c360I** denote conventional images, 360° images, 360° videos, and cross-frame correlated 360° images, respectively. **Res.** refers to image resolution. **Temp. (#)** indicates template-based annotations, with the number of templates in parentheses. **Auto** refers to annotations generated automatically using vision models. **Manual** denotes human-crafted annotations. **#Samples (#)** indicates the number of unique samples, with test samples shown in parentheses. **Avail.** indicates whether the dataset is publicly accessible at submission.

Dataset	Type	Res.	Scene	Annotation	#Samples	#Tasks	FP-IR	FP-IC	PP-IR	PP-IC	SR-Os	SR-OV	DG	Avail.
HR-Bench [36]	CI	8K	Diverse	Manual	200 (200)	2	✓	✗	✗	✗	✓	✗	✗	✗
Pano-AVQA [45]	360V	N/A	Diverse	Temp. (7)	51.7K (5.3K)	2	✗	✗	✗	✗	✓	✗	✗	✗
CFpano [47]	c360I	1K	Indoor	Temp. (36)	8094 (1619)	17	✗	✗	✗	✗	✗	✗	✗	✗
VQA360 [8]	360I	1K	Indoor	Temp. (17)	16,945 (6962)	3	✗	✗	✗	✗	✓	✓	✗	✗
OmniVQA [48]	360I	1K	Indoor	Temp. (6)	5652 (800)	2	✗	✗	✓	✗	✓	✗	✗	✗
ODI-Bench [42]	360I	Mixed	Diverse	Auto&Manual	4254 (4254)	10	✗	✗	✓	✗	✓	✓	✗	✗
360Bench (Ours)	360I	7K	Diverse	Manual	1532 (1532)	7	✓	✓	✓	✓	✓	✓	✓	✓

To bridge this gap, we introduce **360Bench**, a 360° VQA benchmark with 7K-resolution images spanning indoor, outdoor, and aerial scenes. 360Bench comprises 1,532 unique samples featuring curated annotations across seven (sub)-tasks. Table 1 compares 360Bench with existing VQA datasets in terms of image resolution, scene diversity, annotation methods, and task scopes. Using 360Bench, we systematically evaluate seven MLLMs and six enhancement methods, revealing notable performance gaps: the best model achieves only 46.5% accuracy, far below human performance of 86.3%. Regarding the projection format, CMP shows a clear advantage on projection-distorted tasks (up to 14.1%), while ERP generally performs better on spatial reasoning tasks (up to 14.5%), underscoring their complementary strengths.

These results highlight the fundamental limitations of existing MLLMs in understanding 360° scenes. While fine-tuning could improve performance, it is computationally expensive, labor-intensive, and must be repeated for each downstream task [13, 25]. Moreover, fine-tuning risks catastrophic forgetting, erasing general knowledge acquired during pretraining [22, 26, 46]. In contrast, training-free approaches offer a scalable and practical alternative that preserves pre-trained knowledge, crucial for large-scale MLLMs.

To address the challenges of 360° image perception, we propose **Free360**, a training-free scene-graph-based framework for high-resolution single-image 360° VQA. The use of scene graphs enables (1) decomposition of the reasoning process into modular steps, (2) application of adaptive spherical image transformations tailored to each step, and (3) seamless integration of the resulting information into a unified graph representation for final answer generation. Specifically, the scene graph generation (SGG) proceeds in four steps: (i) detecting question-relevant entities as nodes, (ii) extracting fine-grained attributes from cropped entity regions, (iii) inferring inter-entity relations via spherical rotations, and (iv) incorporating spatial relations between entities and the viewer. The resulting scene graph is then serialized into structured text and fed to the MLLM for question answering. Experiments show that Free360 substantially improves the

performance of its base MLLM, yielding gains of up to 22.9% on subtasks and 7.3% overall, showing its effectiveness in bridging the gap between human and model perception of 360° scenes.

Our contributions are summarized as follows:

- We introduce 360Bench, a high-resolution benchmark for single-image 360° VQA, comprising seven subtasks designed to evaluate fine-grained perception, spatial reasoning, and robustness to projection-induced challenges.
- We systematically evaluate thirteen models, including MLLMs and enhanced variants, under both ERP and CMP settings on 360Bench, revealing their limitations in 360° image perception.
- We propose Free360, a training-free scene-graph-based framework for high-resolution single-image 360° VQA that incorporates 360°-specific operations, including entity-centered rotation and view mapping, and consistently improves the performance of its base MLLM.

2 Related Work

2.1 360° VQA Datasets

Several benchmarks have been developed for question answering and reasoning on omnidirectional data [8, 42, 45, 47, 48]. VQA360 [8] focuses on spatial reasoning between objects and between objects and the viewer in 360° images. Pano-AVQA [45] studies grounded question answering in 360° videos. OmniVQA [48] extends 360° image VQA to indoor ERP scenes with tasks centered on object recognition and spatial reasoning under geometric distortion.

Related panoramic reasoning settings have also been explored beyond single-image VQA. For example, CFpano [47] considers cross-frame correlated panoramas, where reasoning is performed over multiple related panoramic observations rather than a single 360° image. In addition, a very recent concurrent benchmark, ODI-Bench [42], broadens the problem from VQA to more general omnidirectional image understanding with a wider set of tasks.

Compared with these efforts, our benchmark is designed specifically for high-resolution single-image 360° VQA with manual annotations and seven subtasks targeting fine-grained perception, spatial reasoning, and robustness to projection-induced challenges. Thus, 360Bench complements existing omnidirectional benchmarks by emphasizing the setting where a single high-resolution panoramic image must support both local perception and global reasoning.

2.2 MLLM Enhancement Methods

Despite recent advances, state-of-the-art MLLMs still show clear limitations in fine-grained perception and complex reasoning. Existing enhancement methods for conventional images, such as SEAL [38], DC^2 [36], and ZoomEye [33], typically rely on patch-wise search or retrieval to identify informative regions and use them to assist downstream reasoning.

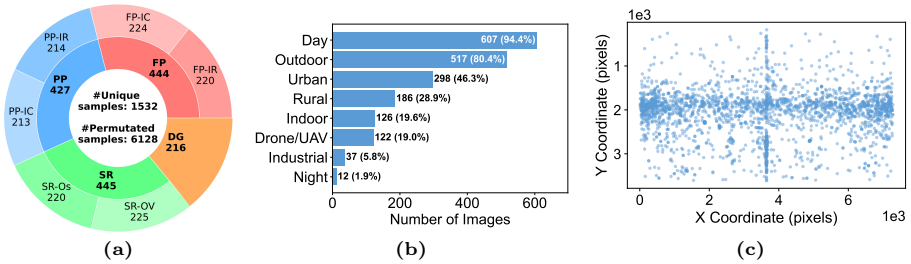


Fig. 2: Statistics of 360Bench: (a) Task breakdown, (b) Image distribution, (c) Spatial distribution of annotated bounding boxes.

For omnidirectional images, dedicated enhancement strategies have only recently begun to emerge. The work in [48] proposes a rule-based reinforcement learning approach to fine-tune Qwen2.5-VL for 360° reasoning. Concurrent work [42] introduces Omni-CoT, a training-free reasoning framework for omnidirectional image understanding. Given a question, Omni-CoT first generates an initial answer from textual descriptions of six perspective views, and subsequently refines the prediction using crops of relevant objects.

Our method is complementary to these efforts. Rather than relying on additional training, we focus on high-resolution single-image 360° VQA and introduce a training-free framework that constructs question-relevant scene graphs through modular reasoning steps. This design enables the use of 360°-specific operations, such as entity-centered rotation and view mapping, to better model fine-grained attributes and spatial relations.

2.3 MLLM-based Scene Graph Generation

Recent studies have employed MLLMs for SGG [5, 10, 41]. To mitigate the substantial cost of manual annotation, image caption datasets are frequently leveraged to automatically generate SGG datasets for training models [18, 21, 34, 43]. Alternatively, [10] introduces a training-free approach in which an LLM-based scene graph parser constructs scene graphs directly from image captions generated by MLLMs. Despite these advances, no prior work has explored using MLLMs for question-relevant scene graph generation, in which graphs are constructed specifically for each question.

3 360Bench Benchmark

3.1 Task Definition

360Bench consists of four task categories: Fine-grained Perception (FP), Projection-distorted Perception (PP), Spatial Reasoning (SR), and Direction-Giving (DG), as summarized in Fig. 2a. Each task targets a distinct aspect of multimodal understanding under panoramic settings.

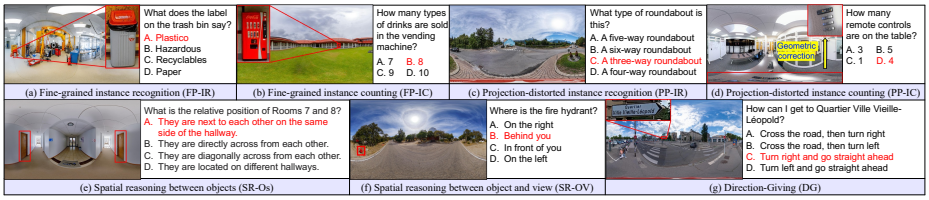


Fig. 3: Illustration of tasks included in 360Bench. Correct answers and relevant entities are highlighted in red. 360° image sources: [12, 28].

The Fine-grained Perception (FP) task evaluates a model's ability to recognize and reason about small or visually subtle instances, such as objects and text, within complex 360° scenes. It includes two subtasks: Instance Recognition (FP-IR), which focuses on identifying instance attributes, as illustrated in Fig. 3(a), and Instance Counting (FP-IC), which measures the accurate enumeration of target objects, as shown in Fig. 3(b).

The Projection-distorted Perception (PP) task assesses a model's robustness to geometric distortion and object fragmentation inherent in 360° projections. It contains two subtasks: Instance Recognition (PP-IR), which queries the attributes of distorted or split objects, and Instance Counting (PP-IC), which evaluates counting accuracy under severe distortions. To construct this task, we selected and annotated images with boundary-fragmented or heavily distorted objects. For instance, Fig. 3(c) shows a PP-IR example where a road divided across the left and right boundaries appears as multiple paths, while Fig. 3(d) depicts a PP-IC example where elongated remote controls complicate counting.

The Spatial Reasoning (SR) task evaluates a model's ability to infer spatial relations in 360° scenes. It comprises 2 subtasks: Object-to-Object (SR-Os) reasoning, which examines relative positioning among objects (see Fig. 3(e)), and Object-to-Viewer (SR-OV) reasoning, which focuses on the spatial relation between objects and the viewer (see Fig. 3(f)).

The Direction-Giving (DG) task tests goal-directed visual reasoning, requiring models to interpret spatial cues and plan multi-step routes toward a target, either visible or indicated by directional signs (see Fig. 3(g)). Beyond SR, DG requires multi-step route planning that relies on holistic scene understanding.

It is noted that for both the SR-OV and DG tasks, the viewer is assumed to be facing the direction corresponding to the center of the ERP image, which serves as the consistent reference viewpoint at longitude 0° and latitude 0°.

3.2 Image Collection

To construct 360Bench, we collected 643 omnidirectional images from Flickr [11], NOIRLab [28], and Insta360 [15]. All images are in ERP format and publicly available for download. Figure 2b shows the distribution across scene types and capture conditions. 360Bench is dominated by daytime, outdoor, urban scenes, but also includes indoor, drone/UAV, and nighttime images, reflecting diverse

real-world scenarios. The image resolutions range from 7296×3648 to 21344×10672; images larger than 7K were downsampled to 7296×3648 for consistency.

3.3 Dataset Annotations

Our 360Bench comprises 1,532 unique samples, each pairing a 360° image with a question and four answer options. All the questions and answer options are carefully designed by human annotators to ensure that correct answers cannot be inferred without truly understanding image content.

Three trained annotators created the annotations following a detailed protocol. For each assigned 360° image, an annotator first explored the scene in Virtual Reality (VR) viewing mode, freely adjusting the viewing direction and zoom level to identify key objects. The objects were then assigned to subtasks based on their characteristics. For each sample, the annotator composed the corresponding question, along with its correct answer and three plausible yet incorrect answers. Bounding boxes were also annotated for the objects relevant to each question. Each annotator spent approximately three weeks completing their assigned set, reflecting the effort required to ensure high-quality annotations. Quality control was enforced through cross-evaluation among annotators to resolve ambiguous cases. This procedure ensures a reliable and challenging benchmark for evaluating MLLMs on 360° image perception. Refer to the supplementary material for further details on the annotation protocol.

On average, each image contributes 2.38 unique samples (1,532 samples across 643 images). Each subtask contains from 213 to 225 unique samples, as shown in Fig. 2a. To mitigate answer position bias, we applied cyclic permutation to the answer options $[36, 50]$, resulting in a total of 6,128 evaluation samples.

Figure 2c shows the spatial distribution of bounding boxes in 360Bench, where each point represents a box center. The annotations cover the full ERP image, with higher density near the equator, reflecting typical object layouts in 360° scenes. Additional statistics are provided in the supplementary material.

4 Proposed Method

Figure 4 shows an overview of our proposed method, **Free360**. Given a question Q and a 360° image I, Free360 answers the question by generating a question-relevant scene graph and reasoning over it. We decompose the SGG process into four explicit steps: (1) Entity Identification, detecting entities relevant to the question Q from the image I; (2) Attribute Extraction, deriving descriptive attributes for each detected entity; (3) Inter-Entity Relation Detection, capturing spatial relations among entities based on spherically rotated images; and (4) Entity-View Relation Detection, modeling spatial relations between entities and camera views (or the viewer).

Each step is guided by a task-specific prompt and a few in-context examples; see the supplementary material for details. Motivated by the complementary strengths of CMP and ERP observed in our analysis (Table 3), we design Free360

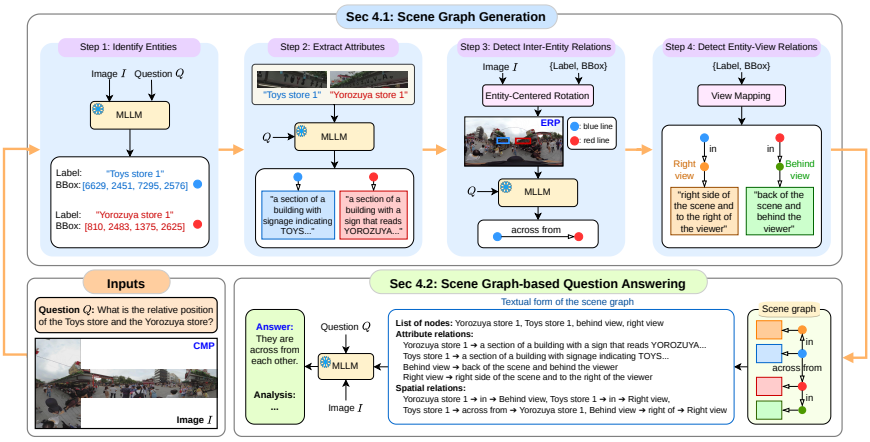


Fig. 4: Overview of Free360. The model performs question answering via scene graph generation through four steps (Sec. 4.1): (1) Entity Identification, detecting entities relevant to the question Q from the CMP image I; (2) Attribute Extraction, deriving descriptive attributes from each entity crop; (3) Inter-Entity Relation Detection, capturing spatial relations among entities; and (4) Entity-View Relation Detection, modeling spatial relations between entities and camera views. The resulting scene graph, represented in textual form, is then fed into the MLLM to generate the final answer and reasoning analysis (Sec. 4.2). 360° image source: [15].

as a hybrid framework that leverages CMP in Step 1 for robust object detection and ERP in Step 3 for spatial reasoning. The resulting scene graph serves as an explicit reasoning scaffold, integrating pertinent information for the MLLM to infer the final answer.

4.1 Scene Graph Generation

The goal of SGG is to construct a scene graph that captures the entities, their attributes, and spatial relations required for reasoning to answer the question Q. We define a scene graph as $G = (\mathcal{N}, \mathcal{R})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{R} \subseteq \mathcal{N} \times \mathcal{N}$ represents the sets of relations between them. Each node $n_i \in \mathcal{N}$ is defined as $n_i = (l_i, a_i)$, where l_i and a_i are the node's label and attribute, respectively. Each directed relation $r_{i,j} \in \mathcal{R}$ represents a semantic relation from node n_i to node n_j.

The graph includes two types of nodes: (1) **entity nodes** n^e, representing visible entities within the image relevant to the question, and (2) **view nodes** n^v, corresponding to the six cube faces of the CMP which represent the six principal directions relative to the viewer (i.e., top, bottom, left, right, front, behind). Accordingly, we define three types of relations: **inter-entity relations** r^{ee}, **entity-view relations** r^{ev}, and **inter-view relations** r^{vv}.

Constructing G with an MLLM is inherently challenging [9, 20]. To make this process tractable, we adopt a structured, step-by-step strategy inspired by the chain-of-thought paradigm [4, 37, 49], consisting of four main steps outlined below.

Step 1: Entity Identification To alleviate geometric distortion, the CMP image I is used as the visual input to Free360. We prompt the MLLM with the CMP image I and question Q to identify and localize entities in the image that are semantically relevant to the question. For each detected entity node n_i^e, the output contains a label l_i^e and its 2D bounding box $(x_{i,1}^e, y_{i,1}^e, x_{i,2}^e, y_{i,2}^e)$, where $(x_{i,1}^e, y_{i,1}^e)$ and $(x_{i,2}^e, y_{i,2}^e)$ denote the top-left and bottom-right corners, respectively. Each entity is subsequently assigned a unique index to differentiate instances sharing the same label (e.g., 'person 1').

Step 2: Attribute Extraction For each entity node n_i^e, we extract its image crop from I using the bounding box $(x_{i,1}^e, y_{i,1}^e, x_{i,2}^e, y_{i,2}^e)$. Cropping not only reduces computational complexity but also allows the model to focus on fine-grained features, such as text, patterns, or logos, that are difficult to discern from the full image. We prompt the MLLM with the label and its associated image crop as inputs to generate a textual attribute a_i^e describing the entity in a manner relevant to answering the question Q. To integrate these attributes into the textual representation of the scene graph, we represent attribute relations as '$l_i^e \rightarrow a_i^e$', linking each entity node to its corresponding attribute.

Step 3: Inter-Entity Relation Detection Reasoning about spatial relations in 360° scenes can benefit from centering relevant entities, analogous to how humans shift their viewpoint. For each entity pair (n_i^e, n_j^e), we generate a rotated ERP image I' that centers the pair, facilitating the MLLM's spatial reasoning. The rotation is parameterized by angles ϕ' (rotation about the y-axis) and θ' (rotation about the z-axis), computed from the bounding boxes $(x_{i,1}^e, y_{i,1}^e, x_{i,2}^e, y_{i,2}^e)$ and $(x_{j,1}^e, y_{j,1}^e, x_{j,2}^e, y_{j,2}^e)$.

Specifically, we define the center $\mathbf{c}^*$ of the entity pair as the midpoint of the minimal spherical bounding box enclosing both entities; refer to our supplementary material for further details. The rotation angles (ϕ', θ') correspond to the longitude and latitude of $\mathbf{c}^*$. We obtain the rotated image I' from I by the rotation matrix

$$\mathbf{R}_{\mathrm{c}} = \begin{bmatrix} \cos\phi'\cos\theta' & -\cos\phi'\sin\theta' & \sin\phi' \\ \sin\theta' & \cos\theta' & 0 \\ -\sin\phi'\cos\theta' & \sin\phi'\sin\theta' & \cos\phi' \end{bmatrix}.$$

We then highlighted the entities in I' with colored bounding boxes and a legend, i.e., "l_i^e`:blue line`,l_j^e`:red line`". We prompt the MLLM to analyze I' along with the legend to describe the spatial relation r_{ij}^{ee} between the entities. Inter-entity relations are then encoded textually as '$l_i^e \rightarrow r_{ij}^{ee} \rightarrow l_j^e$.'

Table 2: Labels and attributes of the six view nodes.

Label	Attribute
`left view`	Left side of the scene and to the left of the viewer.
`right view`	Right side of the scene and to the right of the viewer.
`front view`	Front of the scene and front of the viewer.
`behind view`	Back of the scene and behind the viewer.
`top view`	Top of the scene and above the viewer.
`bottom view`	Bottom of the scene and below the viewer.

Step 4: Entity-View Relation Detection We define six view nodes $\mathcal{N}^v = \{n_q^v\}_{q=1}^6$ corresponding to the cube faces in the CMP format; see Table 2. Inter-view relations r^{vv} are predefined and incorporated into the scene graph to capture spatial dependencies between views. For instance, the relation between `front view` and `behind view` is represented as '`front view` → `opposite` → `behind view`' since they lie on opposite cube faces.

To assign entities to views, we define a mapping function $f : \mathbb{R}^2 \to \mathcal{N}^v$ that maps 2D pixel coordinates in the CMP image to the corresponding view node (see our supplementary material for more details). For each entity node n_i^e, we compute the center of its bounding box $(\bar{x}_i, \bar{y}_i)$, and determine its associated view node $n_q^v = f(\bar{x}_i, \bar{y}_i)$. The textual format of entity-view relation r_{iq}^{ev} is then represented as '$l_i^e \to \texttt{in} \to l_q^v$', linking each entity node to its corresponding view node in the scene graph.

4.2 Scene Graph-Based Question Answering

The constructed scene graph from previous SGG steps is serialized into a structured textual form and fed to the MLLM for reasoning. Specifically, we represent a scene graph with a list of relevant nodes (detected entities and views) and their spatial and attribute relations as follows:

$$\text{List of nodes: } \begin{cases} \{l_i^e\}_{i=1}^M, \\ \{l_q^v\}_{q=1}^K \end{cases}$$

$$\text{Attribute relations: } \begin{cases} \{l_i^e \to a_i^e\}_{i=1}^M, \\ \{l_q^v \to a_q^v\}_{q=1}^K \end{cases}$$

$$\text{Spatial relations: } \begin{cases} \{l_i^e \to r_{iq}^{ev} \to l_q^v\}_{i=1}^M, \\ \{l_i^e \to r_{ij}^{ee} \to l_j^e\}_{1 \le i < j \le M}, \\ \{l_q^v \to r_{qp}^{vv} \to l_p^v\}_{1 \le q < p \le K}. \end{cases}$$

Here, M and K denote the numbers of entities and views, respectively. Figure 5 shows a serialized scene graph example.

To generate an answer, we prompt the MLLM with instructions to produce the reasoning analysis and select the most plausible option, along with the serialized graph text, the question, and its answer options. To ensure answer reliability, the MLLM is instructed to output '`CANNOT ANSWER`' if the provided graph lacks

List of nodes: Yorozuya store 1, Toys store 1, Behind view, Right view

Attribute relations:

- Yorozuya store 1→ a section of a building with a sign that reads YOROZUYA...
- Toys store 1→ a section of a building with signage indicating TOYS...
- Behind view → back of the scene and behind the viewer
- Right view → right side of the scene and to the right of the viewer

Spatial relations:

- Yorozuya store 1 → in → Behind view
- Toys store 1 → in → Right view
- Toys store 1 → across from → Yorozuya store 1
- Behind view → right of → Right view

Fig. 5: Example of a scene graph in the serialized textual form.

sufficient information. In such cases, the image I is used as a fallback input to the MLLM for generating the final answer.

5 Experiments

5.1 Human Evaluation

We conducted a subjective study to evaluate human performance on 360Bench. Five participants (university students and staff aged 22–34) took part in the study. We developed a web-based user interface using A-Frame [1] to present 360° images in VR mode to the participants. The participants could freely adjust their viewing direction to explore each 360° scene using a mouse. To mitigate potential motion sickness from prolonged 360° viewing, each participant answered 70 randomly sampled questions (10 per subtask) from 360Bench, taking approximately 20–50 minutes. Refer to our supplementary material for further details.

5.2 Experimental Settings

Evaluated models. We evaluate seven representative open-source and proprietary MLLMs on 360Bench: GPT-4o [29], Gemini Flash 2.5 [35], Gemini Pro 2.5 [35], Qwen2.5-VL-7B [2], DeepSeek-VL2-Small [39], LLaVA-v1.5-7B [24], and LLaVA-CoT [40]. We also evaluate six enhancement methods: SEAL [38], DC^2 [36], ZoomEye [33], VisCoT-7B-336 [32], DeepEyes [51], and Omni-CoT [42]. A full evaluation of 27 models is provided in the supplementary material.

Although several other 360° VQA models have been proposed [8, 48], they are excluded from our evaluation because (1) their required large-scale training datasets are unavailable, (2) their pretrained weights have not been released, and (3) models such as [8] are not compatible with MLLM-based inference.

Evaluation metrics. We report accuracy as the primary evaluation metric, calculated as the proportion of correctly answered samples. Except for proprietary models whose inference time is measured via API calls, we also report the

average per-sample inference time on an NVIDIA H200 GPU (in seconds) to assess computational efficiency.

Implementation details. All models are configured to generate outputs deterministically using greedy decoding to ensure reproducibility. We convert ERP images into CMP format using 360Lib [17]. The CMP resolution is set to 7296×5472, matching the equatorial pixel count of the ERP format to ensure a fair comparison. Regarding image processing, proprietary models directly handle images via their APIs. For other models, images are preprocessed according to their default settings to ensure evaluation under their intended operating conditions. Specifically, Qwen2.5-VL-7B and its enhancement methods process ERP images at 5068×2520 and CMP images at 4116×3080. DeepSeek-VL2-Small uses dynamic tiling, with tile sizes of 384 and 1024. The remaining models resize inputs to their maximum supported resolutions—224 for SEAL and 336 for the others.

For the enhancement methods, SEAL, VisCoT-7B-336, and DeepEyes are tested using their official fine-tuned models, while the training-free methods DC^2, ZoomEye, and Omni-CoT adopt Qwen2.5-VL-7B as the base MLLM. Since Omni-CoT's source code is unavailable at submission, we re-implement it based on the original paper. For a fair comparison with other training-free methods, we also employ Qwen2.5-VL-7B as its base model for Free360.

5.3 Results on 360Bench

Table 3 compares human and model performance on our 360Bench; the supplementary material presents further qualitative results. Human participants achieve the highest accuracy of 86.3% by immersing themselves in 360° images and answering each question in 28.9 seconds on average. They perform strongly on recognition, spatial reasoning, and direction-giving tasks ($\geq$84%) but show lower accuracy on counting (82% for FP-IC and 78% for PP-IC), highlighting the inherent counting difficulty in complex 360° scenes.

State-of-the-art MLLMs, including GPT-4o and Gemini Pro 2.5, still struggle with omnidirectional perception. While Gemini Pro 2.5 achieves the best overall accuracy (46.5%), it falls far short of human-level accuracy (86.3%). Furthermore, its black-box nature makes direct comparisons with open-source models less informative. Among open-source models, Qwen2.5-VL-7B leads with 38.1%, underscoring the persistent gap between current MLLMs and humans.

Projection format strongly influences MLLM's performance. Overall, CMP achieves higher accuracy than ERP across most models (up to +2.6% for Gemini Pro 2.5). CMP shows particularly strong gains on projection-distorted tasks (e.g., +14.1% on PP-IR with LLaVA-CoT and +12.2% on PP-IC with Gemini 2.5 Flash). In contrast, ERP generally performs better on spatial reasoning tasks (e.g., +2.4% on SR-Os for LLaVA-CoT and +14.5% on SR-OV for Gemini 2.5 Flash, and +6.8% on DG for Gemini 2.5 Pro). These results reveal complementary strengths between the two formats: CMP is more effective for object perception under projection distortion, whereas ERP benefits spatial reasoning.

Among enhancement methods, ZoomEye improves fine-grained perception on FP-IR (+1.9%) but weakens holistic reasoning on FR-IC (-1.3%) over its

Table 3: Human and model evaluation results. The highest score for each task is highlighted in bold. *Acc.* denotes per-task accuracy; *Difference* indicates the difference between Free360 and Qwen2.5-VL-7B in accuracy and inference time.

Model	Projection format	FP (Acc.)			PP (Acc.)			SR (Acc.)			DG (Acc.)	Overall (Acc.)	Inf. Time (seconds)
		IR	IC	Avg.	IR	IC	Avg.	Os	OV	Avg.			
Random guess	—	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	—
Human	—	96.0	82.0	89.0	86.0	78.0	82.0	84.0	90.0	87.0	88.0	86.3	28.9
MLLMs													
GPT-4o	ERP	48.6	32.0	40.3	54.2	38.8	46.5	37.7	35.7	36.7	38.5	40.7	1.5
	CMP	45.5	30.8	38.1	**63.0**	44.0	**53.5**	37.4	34.0	35.7	35.3	41.3	1.6
Gemini Flash 2.5	ERP	62.3	40.0	51.0	44.5	33.8	39.2	37.3	40.3	38.8	40.6	42.7	10.8
	CMP	57.5	41.9	49.6	56.5	46.0	51.3	37.1	25.8	31.5	34.7	42.6	12.5
Gemini Pro 2.5	ERP	**63.2**	43.5	**53.3**	41.1	34.7	37.9	38.0	**42.9**	40.5	**44.0**	43.9	18.4
	CMP	61.5	**44.8**	53.0	52.2	**46.7**	49.5	**46.8**	36.9	**41.9**	37.2	**46.5**	18.3
Qwen2.5-VL-7B	ERP	59.5	27.9	43.6	42.2	36.0	39.1	30.7	31.2	31.0	33.7	37.3	2.1
	CMP	58.1	27.0	42.4	50.6	42.0	46.3	28.4	30.4	29.4	30.6	38.1	2.2
DeepSeek-VL2-Small	ERP	41.8	22.7	32.2	42.4	39.3	40.9	27.2	32.6	29.9	32.4	33.9	0.9
	CMP	40.3	19.2	29.7	55.4	43.1	49.2	28.2	29.7	28.9	31.1	35.1	0.9
LLaVA-v1.5-7B	ERP	30.7	8.4	19.4	29.3	35.9	32.6	22.4	26.7	24.5	26.6	25.6	**0.4**
	CMP	32.5	9.3	20.8	33.4	39.6	36.5	23.5	23.9	23.7	26.3	26.8	0.5
LLaVA-CoT	ERP	34.9	16.7	25.7	41.5	33.3	37.4	27.2	30.0	28.6	27.5	30.1	2.7
	CMP	33.3	16.9	25.0	55.6	37.3	46.5	24.8	30.9	27.8	27.6	32.2	3.0
MLLM Enhancement Models													
SEAL	ERP	31.3	24.6	27.9	30.4	27.8	29.1	23.6	20.4	22.0	18.9	25.2	14.6
	CMP	33.6	21.4	27.5	32.1	30.4	31.3	23.6	18.7	21.1	20.7	25.7	14.6
DC^2	ERP	40.0	16.0	27.9	44.0	36.4	40.2	22.0	25.1	23.6	21.5	29.1	617.9
	CMP	39.7	11.0	25.2	49.1	34.9	42.0	20.2	23.1	21.6	20.2	28.1	761.6
ZoomEye	ERP	**61.5**	26.6	43.9	44.2	31.9	38.1	25.3	31.3	28.3	28.5	35.5	19.5
	CMP	59.2	22.8	40.8	51.2	30.3	40.7	27.3	29.1	28.2	28.0	35.3	23.9
VisCoT-7B-336	ERP	28.1	23.3	25.7	30.6	35.9	33.3	22.3	26.7	24.4	23.8	27.2	1.1
	CMP	28.3	23.7	26.0	34.8	36.4	35.6	24.6	24.4	24.5	24.2	28.0	**0.8**
DeepEyes	ERP	53.6	25.2	39.3	33.3	32.0	32.7	28.2	26.9	27.5	29.1	32.6	78.4
	CMP	52.6	25.6	39.0	43.5	42.5	43.0	31.4	26.3	28.9	27.3	35.5	66.9
Omni-CoT	Hybrid	58.9	24.3	41.4	49.5	40.3	44.9	28.3	41.0	34.7	35.0	39.5	16.6
Free360 (Ours)	Hybrid	60.7	**30.0**	**45.2**	**54.7**	**43.9**	**49.3**	**33.6**	**53.3**	**43.6**	**41.4**	**45.3**	22.5
Difference w.r.t	ERP	+ 1.1	+ 2.1	+ 1.6	+ 12.5	+ 7.9	+ 10.2	+ 3.0	+ 22.1	+ 12.6	+ 7.8	+ 8.1	+ 20.3
Qwen2.5-VL-7B	CMP	+ 2.6	+ 3.0	+ 2.8	+ 4.1	+ 1.9	+ 3.0	+ 5.2	+ 22.9	+ 14.2	+ 10.9	+ 7.3	+ 18.1

base MLLM (Qwen2.5-VL-7B), indicating that its patch-wise search favors local detail over global understanding. For Omni-CoT, gains of +10.6% on SR-OV and +4.4% on DG suggest that providing explicit view-level information effectively strengthens viewer-centric spatial reasoning. However, Omni-CoT shows marginal declines on most other tasks (up to -2.4%), implying that relying solely on view-level decomposition and object-level cropping may be insufficient for more complex and holistic reasoning.

In contrast, **Free360** achieves the best overall accuracy of 45.3%, outperforming Qwen2.5-VL-7B with CMP by +7.3% and consistently improving performance across all tasks (up to +22.9% on SR-OV). These results demonstrate that scene graph-based reasoning, empowered by 360°-specific operations, enables more robust and comprehensive understanding in omnidirectional settings. These improvements are achieved with a moderate increase in inference time (from 2.1 to 22.5 seconds). Nevertheless, the resulting latency is still within the range of human response time (28.9 seconds). It is comparable to ZoomEye (19.5–23.9 seconds) and slightly higher than Omni-CoT (16.6 seconds), while

Table 4: Ablation study of Free360.

(a) Impact of SGG components

Setting	Acc.	Inf. Time
Full	**45.3**	22.5
w/o Crop	42.7	23.8
w/o Rotate	44.7	**21.0**
w/o EVR	42.8	21.1

(b) Effectiveness across model scales

Model scale	Qwen2.5-VL		Free360	Acc. Gain	
	ERP	CMP	Hybrid	ERP	CMP
3B	35.2	37.2	40.8	+ 5.5	+ 3.5
7B	37.3	38.1	**45.3**	**+ 8.1**	**+ 7.3**
32B	**37.8**	**39.7**	45.0	+ 7.2	+ 5.3

being substantially faster than DeepEyes (66.9–78.4 seconds) and DC^2 (617.9–761.6 seconds). Further discussion on inference time and potential strategies for efficiency improvement is provided in the supplementary material.

5.4 Ablation Study

We conduct an ablation study to evaluate the contribution of each component in Free360 and its effectiveness across different model scales, as summarized in Table 4. Specifically, *Full* denotes the complete model; *w/o Crop* uses the full image instead of image crops for the attribute extraction; *w/o Rotate* denotes removing the spherical rotation in the inter-entity relation detection; and *w/o EVR* excludes the entity-view relation detection step. Under the *Full* setting, we further assess the effectiveness of Free360 across varying base MLLM scales (i.e., 3B, 7B, and 32B).

Table 4(a) isolates the contribution of each design choice in Free360. Using image cropping for attribute extraction yields a 2.7% accuracy gain and reduces inference time by 1.4 seconds, suggesting improved focus and efficiency. Spherical rotation enhances inter-entity relation detection, adding 0.7% accuracy at a 1.5 second inference cost. Finally, modeling entity-view relations provides a further 2.6% improvement with an overhead of 1.4 seconds, confirming its complementary benefit.

Table 4(b) reports the performance of the base MLLM (Qwen2.5-VL) and our Free360 under the Full setting across different model scales. Free360 consistently improves performance across all base model sizes. The largest gain is observed at 7B, where Free360 achieves 45.3% accuracy, yielding improvements of +8.1% over ERP and +7.3% over CMP. For 3B and 32B, the improvements range from +3.5% to +7.2%. These results confirm the general effectiveness of Free360 across varying model scales.

6 Conclusion and Future Work

In this paper, we have presented a comprehensive study on 360° image perception with MLLMs through 360Bench, a 360° VQA benchmark. Using 360Bench, we have systematically evaluated thirteen state-of-the-art models, including both MLLMs and enhanced variants, revealing their limitations in reasoning over panoramic scenes. To address this, we have introduced Free360, a training-free, scene-graph-based framework for high-resolution 360° VQA, empowered

by 360-specific operations, including entity-centered rotation and view mapping. Free360 consistently improves its base MLLM across all tasks, achieving a 7.3% overall improvement while remaining computationally efficient.

Future work on training-free methods using scene graphs for 360° tasks may explore finer perception modules to strengthen local reasoning, such as patch-wise retrieval. Extending Free360 for 360° video understanding is also a promising direction. We hope this study provides a foundation for advancing multimodal reasoning in omnidirectional environments and inspires further development of robust, generalizable MLLMs for 360° perception.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP23H00482 and JP26K21241. The authors gratefully acknowledge hapePHOTOGRAPHIX and Ozymandias for kindly granting permission to use their panoramic photographs from Flickr in this study. Full credit is given to both photographers, with additional credit to hapePHOTOGRAPHIX via his official website[3].

References

1. AframeVR: A-Frame: A web framework for building browser based 3D, AR and VR experiences. `https://github.com/aframevr/aframe` (2025), accessed: 2025-11-05
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025), `https://arxiv.org/abs/2502.13923`
3. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. ACM transactions on intelligent systems and technology **15**(3), 1–45 (2024)
4. Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., Che, W.: Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567 (2025)
5. Chen, Z., Wu, J., Lei, Z., Zhang, Z., Chen, C.: Gpt4sgg: Synthesizing scene graphs from holistic and region-specific narratives. arXiv preprint arXiv:2312.04314 (2023)
6. Cheng, H.T., Chao, C.H., Dong, J.D., Wen, H.K., Liu, T.L., Sun, M.: Cube padding for weakly-supervised saliency prediction in 360 videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1420–1429 (2018)
7. Chiang, J.C., Yang, C.Y., Dedhia, B., Char, Y.F.: Saliency-driven rate-distortion optimization for 360-degree image coding. Multimedia Tools and Applications **80**(6), 8309–8329 (2021)
8. Chou, S.H., Chao, W.L., Lai, W.S., Sun, M., Yang, M.H.: Visual question answering on 360° images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1607–1616 (2020)

[3] https://www.hapePHOTOGRAPHIX.com

9. Dutta, A., Mehrab, K.S., Sawhney, M., Neog, A., Khurana, M., Fatemi, S., Pradhan, A., Maruf, M., Lourentzou, I., Daw, A., et al.: Open World Scene Graph Generation using Vision Language Models. arXiv preprint arXiv:2506.08189 (2025)
10. Elskhawy, A., Li, M., Navab, N., Busam, B.: Prism-0: A predicate-rich scene graph generation framework for zero-shot open-vocabulary tasks. arXiv preprint arXiv:2504.00844 (2025)
11. Flickr: https://www.flickr.com/ (2025), accessed: 2025-06-20
12. hapePHOTOGRAPHIX: 360° Image. https://www.flickr.com/photos/hapephotographix/ (2022), credit: hapePHOTOGRAPHIX (https://www.hapePHOTOGRAPHIX.com). Accessed: 2025-07-08
13. He, Y., Luo, X.: Tensor low-rank orthogonal compression for convolutional neural networks. IEEE/CAA Journal of Automatica Sinica **13**(1), 227–229 (2026)
14. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., Wei, F.: Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045 (2023), https://arxiv.org/abs/2302.14045
15. Insta360: Professional 360 VR Camera. https://www.insta360.com/product/insta360-pro2 (2025), accessed: 2025-06-20
16. Jung, D., Choi, J., Lee, Y., Jeong, S., Lee, T., Manocha, D., Yeon, S.: Edm: Equirectangular projection-oriented dense kernelized feature matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6337–6347 (2025)
17. JVET: Algorithm Descriptions of Projection Format Conversion and Video Quality Metrics in 360Lib Version 5 (2017)
18. Kim, K., Yoon, K., Jeon, J., In, Y., Moon, J., Kim, D., Park, C.: Llm4sgg: Large language models for weakly supervised scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28306–28316 (2024)
19. Li, M., Chen, K., Bi, Z., Liu, M., Peng, B., Niu, Q., Liu, J., Wang, J., Zhang, S., Pan, X., Xu, J., Feng, P.: Surveying the mllm landscape: A meta-review of current surveys. arXiv preprint arXiv:2409.18991 (2024), https://arxiv.org/abs/2409.18991
20. Li, R., Zhang, S., Lin, D., Chen, K., He, X.: From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28076–28086 (2024)
21. Li, X., Chen, L., Ma, W., Yang, Y., Xiao, J.: Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4204–4213 (2022)
22. Liao, C., Xie, R., Sun, X., Sun, H., Kang, Z.: Exploring forgetting in large language model pre-training. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2112–2127 (2025)
23. Lin, J., Men, R., Yang, A., Zhou, C., Ding, M., Zhang, Y., Wang, P., Wang, A., Jiang, L., Jia, X., Zhang, J., Zhang, J., Zou, X., Li, Z., Deng, X., Liu, J., Xue, J., Zhou, H., Ma, J., Yu, J., Li, Y., Lin, W., Zhou, J., Tang, J., Yang, H.: M6: A chinese multimodal pretrainer. arXiv preprint arXiv:2103.00823 (2021), https://arxiv.org/abs/2103.00823
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2024), https://arxiv.org/abs/2310.03744

25. Long, S., Wang, L., Zhao, Z., Tan, Z., Wu, Y., Wang, S., Wang, J.: Training-free unsupervised prompt for vision-language models. arXiv preprint arXiv:2404.16339 (2024), https://arxiv.org/abs/2404.16339
26. Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., Zhang, Y.: An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint arXiv:2308.08747 (2025), https://arxiv.org/abs/2308.08747
27. Nguyen, D.V., Tran, H.T., Pham, A.T., Thang, T.C.: An optimal tile-based approach for viewport-adaptive 360-degree video streaming. IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**(1), 29–42 (2019)
28. NOIRLab: 360 Panorama. https://noirlab.edu/public/images/archive/category/360pano/ (2025), accessed: 2025-06-20
29. OpenAI: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024), https://arxiv.org/abs/2410.21276
30. Prasad, P., Balse, R., Balchandani, D.: Exploring multimodal generative ai for education through co-design workshops with students. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3706598.3714146, https://doi.org/10.1145/3706598.3714146
31. Rao, S., Kumar, V., Kifer, D., Giles, C.L., Mali, A.: Omnilayout: Room layout reconstruction from indoor spherical panoramas. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3706–3715 (2021)
32. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Track on Datasets and Benchmarks (2024)
33. Shen, H., Zhao, K., Zhao, T., Xu, R., Zhang, Z., Zhu, M., Yin, J.: Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. arXiv preprint arXiv:2411.16044 (2024)
34. Shi, J., Zhong, Y., Xu, N., Li, Y., Xu, C.: A simple baseline for weakly-supervised scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16393–16402 (2021)
35. Team, G.: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv preprint arXiv:2507.06261 (2025), https://arxiv.org/abs/2507.06261
36. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. Proceedings of the AAAI Conference on Artificial Intelligence **39**(8), 7907–7915 (2025)
37. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Proceedings of the 36th International Conference on Neural Information Processing Systems (2022)
38. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024)
39. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., Ruan, C.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024), https://arxiv.org/abs/2412.10302

40. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025)
41. Xu, M., Wu, M., Zhao, Y., Li, J.C.L., Ou, W.: Llava-spacesgg: Visual instruct tuning for open-vocabulary scene graph generation with enhanced spatial relations. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6362–6372. IEEE (2025)
42. Yang, L., Duan, H., Tao, R., Cheng, J., Wu, S., Li, Y., Liu, J., Min, X., Zhai, G.: Odi-bench: Can mllms understand immersive omnidirectional environments? (2025), https://arxiv.org/abs/2510.11549
43. Ye, K., Kovashka, A.: Linguistic structures as weak supervision for visual scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8289–8299 (2021)
44. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12) (2024)
45. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-AVQA: Grounded Audio-Visual Question Answering on 360° Videos . In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2011–2021. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00204, https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00204
46. Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y.J., Ma, Y.: Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In: Conference on Parsimony and Learning. pp. 202–227 (2024)
47. Zhang, X., Fu, T., Xu, Z.: Omnidirectional spatial modeling from correlated panoramas. In: Proceedings of the 7th ACM International Conference on Multimedia in Asia. MMAsia '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3743093.3770977, https://doi.org/10.1145/3743093.3770977
48. Zhang, X., Ye, Z., Zheng, X.: Towards Omnidirectional Reasoning with 360-R1: A Dataset, Benchmark, and GRPO-based Method. arXiv preprint arXiv:2505.14197 (2025)
49. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
50. Zheng, C., Zhou, H., Meng, F., Zhou, J., Huang, M.: Large language models are not robust multiple choice selectors. In: The International Conference on Learning Representations (2024), https://openreview.net/forum?id=shr9PXz7T0
51. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025), https://arxiv.org/abs/2505.14362
52. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv preprint arXiv:2504.10479 (2025), https://arxiv.org/abs/2504.10479