Seek to Segment: Active Perception for Panoramic Referring Segmentation

Song Tang[1], Shuming Hu[1], Xincheng Shuai[1], Henghui Ding[1] ✉, and Yu-Gang Jiang[2] ✉

[1] Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, China
[2] Institute of Trustworthy Embodied AI, Fudan University, China
https://henghuiding.com/APRS/

Abstract. Existing referring image segmentation (RIS) models passively process static images captured from fixed perspectives, limiting their applicability in Embodied AI, where agents must perform active perception in the continuous 360° environments. To bridge this gap, we introduce a novel task: **Active Panoramic Referring Segmentation (APRS)**. In this setting, an agent is required to adjust its viewing direction $(\Delta\theta, \Delta\phi)$ to explore the 360° environment, seeking the object specified by a user instruction for segmentation. To tackle this challenging task, we propose **PanoSeeker**, a memory-augmented agent for efficient APRS. Rather than relying on heuristic scanning, PanoSeeker integrates a Vision-Language Model (VLM) with **EgoSphere**—an explicit spatial visual memory. By progressively integrating sequential local observations into a unified 360° representation, EgoSphere enables the agent to plan efficient and non-redundant search trajectories. Once the target is found, the agent performs active viewpoint alignment and outputs the segmentation mask. Furthermore, we curate an expert-annotated search trajectory dataset with memory timelines for Supervised Fine-Tuning, followed by Reinforcement Learning post-training to explicitly optimize PanoSeeker's exploration efficiency. Extensive experiments on our newly established APRS benchmark demonstrate that PanoSeeker achieves superior search efficiency and segmentation accuracy, significantly outperforming adapted state-of-the-art baselines.

Keywords: Agent Memory · Embodied AI · Vision-Language Agent · Active Perception · Referring Image Segmentation · Panoramic Vision

1 Introduction

Referring Image Segmentation (RIS) [1–7] has achieved remarkable progress by segmenting target objects guided by natural language expressions. Powered by recent advancements in Vision-Language Models (VLMs) [8–11], contemporary

✉ Corresponding authors (hhding@fudan.edu.cn, ygj@fudan.edu.cn)

Fig. 1: Active Panoramic Referring Segmentation (APRS) requires an agent to search for and segment a target (e.g., finding the "floor cabinet ") within a continuous 360° environment based on a user instruction. The bottom row illustrates a search trajectory: the agent adjusts its camera pose (θ, ϕ) to reason about spatial relations (e.g., "opposite the bed") and search the target for segmentation.

RIS models [12–16] demonstrate exceptional reasoning ability for precise pixel-level segmentation. However, these methods passively process static images captured from fixed perspectives, which significantly limits their deployment in real-world Embodied AI [17,18], where agents must actively perceive their surroundings to find targets within continuous 360° environments.

Transitioning from passive fixed-view observation to active multi-view perception presents fundamental challenges. In 360° environments, identifying a target necessitates reasoning over complex cross-view spatial relationships (e.g., "the floor cabinet opposite the bed") and performing active exploration. To address these challenges, we introduce a novel task: **Active Panoramic Referring Segmentation (APRS)**. In this setting, an agent is required to adjust its viewing direction $(\Delta\theta, \Delta\phi)$ to explore the 360° environment, seeking the object specified by a user instruction for segmentation, as shown in Fig. 1.

However, applying existing perception methods to this task faces a dilemma. On the one hand, explicit 3D modeling (e.g., point clouds [19–25], or Gaussian Splatting [26–28]) resolves spatial queries by first reconstructing the entire 3D scene, which provides detailed spatial layouts but entails substantial computational and hardware costs. On the other hand, conventional 2D methods [29–31]

that directly process flattened panoramic images introduce severe geometric distortions and wraparound boundaries, while those partitioning the 360° scene into an isolated sequence of perspective views sacrifice global spatial context.

To address this dilemma, we construct the **APRS benchmark**, which leverages abundant and easily accessible panoramic images to simulate realistic 360° egocentric exploration scenes. It eliminates the need for resource-intensive 3D reconstructions while covering diverse indoor and outdoor environments. Moreover, this benchmark features four distinct types of spatial referring expressions (EGO, UNIQ, ALLO, and MULTIHOP) to comprehensively evaluate agents' cross-view spatial reasoning ability across various difficulties. We further provide comprehensive evaluation protocols for benchmarking.

Additionally, we propose **PanoSeeker**, a Vision-Language Agent designed for APRS. To reduce redundant exploration and maintain global spatial context, we design **EgoSphere**, an explicit spatial-visual memory, which integrates sequential local observations into a unified 360° equirectangular panorama (ERP) map. Augmented by EgoSphere, PanoSeeker can perform efficient searching. Once the target is found, the agent performs active viewpoint alignment and generates a segmentation mask. Furthermore, we curate an expert-annotated search trajectory dataset with memory timelines for Supervised Fine-Tuning (SFT). This is followed by Reinforcement Learning (RL) using GRPO [32], which incorporates efficiency and terminal rewards to explicitly optimize PanoSeeker's exploration efficiency. Our main contributions are summarized as follows:

- We introduce Active Panoramic Referring Segmentation (**APRS**), a novel task that requires an agent to actively explore 360° environments to seek and segment targets based on language instructions. To support this, we construct the APRS benchmark, which incorporates four distinct types of spatial referring expressions and comprehensive evaluation protocols.
- We propose **PanoSeeker**, a memory-augmented agent that leverages an explicit spatial visual memory **EgoSphere** to facilitate efficient exploration.
- We contribute an expert-annotated search trajectory dataset with memory timelines for SFT. Additionally, we employ RL with efficiency and terminal rewards to explicitly optimize PanoSeeker's exploration efficiency.
- Extensive experiments demonstrate that PanoSeeker outperforms existing state-of-the-art baselines adapted for this task.

2 Related Work

Referring Image Segmentation. Referring Image Segmentation (RIS) [1,3,4] targets the segmentation of specific objects in an image according to a given textual description. Early approaches [33–35] rely on CNNs to obtain visual features and LSTMs to encode the linguistic expressions, and then combine the two modalities via concatenation or other straightforward fusion operations to form a joint representation. VLT [4,7] is the first to bring the transformer into this task, recasting RIS as an attention problem in which language features serve as queries

over the visual features and the segmentation result is obtained by decoding the transformer output. Liu *et al.* [5] further extend the task to Generalized Referring Expression Segmentation (GRES), which accounts for both multi-target and empty-target cases. More recently, Large Language Models (LLMs) and Vision Language Models (VLMs) [8,36–39] have substantially advanced vision-language tasks through their strong common-sense reasoning ability, offering new opportunities for RIS. Along this line, LISA [12] introduces the [SEG] token, which allows the model to handle expressions that demand complex reasoning and common-sense knowledge. READ [14] treats similarity as reference points to direct where MLLMs should attend during interactive reasoning. Seg-Zero [40] adopts a purely reinforcement-learning scheme (GRPO [32]) to acquire the reasoning process from scratch, producing a reasoning chain prior to the final mask. VisionReasoner [15] presents a unified framework that addresses multiple visual perception tasks through reasoning within a single shared model. Unlike these methods, which depend on static images captured from a fixed viewpoint, our PanoSeeker is able to actively perceive the continuous 360° environment.

Agent Memory. Existing agent memory systems can be broadly categorized into flat-context and structured designs. MemGPT [41] and MemoryOS [42] adopt flat-context paradigms, which leverage paging or streaming mechanisms to extend memory capacity; however, their reliance on raw dialogue logs often introduces significant information redundancy and prohibitive retrieval overhead as histories expand. To enhance semantic coherence and navigability, MemoryBank [43] and A-Mem [44] adopt the structured design to organize memory into hierarchical or graph-based representations. More recently, frameworks such as xMemory [45] have refined these structures by tailoring hierarchical retrieval to the dynamic nature of agentic workloads. Extending beyond textual modalities, vision-centric frameworks like VisMem [46] incorporate short- and long-term latent visual memory into VLMs to enable robust long-term temporal reasoning. Despite these advancements, existing history-based methods face a critical scalability bottleneck, as context length and query latency scale sharply with interaction history. To address this, we propose EgoSphere, a spatial visual memory that maps sequential observations onto a fixed-resolution 360° panoramic canvas; this ensures a constant-length context regardless of the trajectory's length.

Panoramic Vision and Embodied Referring. Panoramic vision provides comprehensive 360° contextual understanding, significantly advancing the ability of intelligent systems in navigation [18, 47] and perception [30, 31, 48]. Early panoramic vision methods primarily focused on static images or videos with a 360° Field-of-View (FoV), addressing challenges such as geometric distortion [49–51], representation learning [48,52–54], and multi-modal integration [30, 31, 55, 56]. However, relying on static, fully observable panoramas diminishes the inherent need for active exploration and spatial reasoning. To address these limitations, recent embodied tasks (e.g., Refer360 [47], REVERIE [18]) integrate linguistic instructions into panoramic environments to inspire models' active perception and spatial reasoning ability. But they remain limited: Refer360 [47] relies on dense, step-by-step instructions that reduce the task to low-level in-

struction following, while REVERIE [18] provides multi-view observations at each step, turning search into a passive selection task. In contrast, our proposed Active Panoramic Referring Segmentation (APRS) requires the agent to adjust its viewing direction to explore the continuous 360° environment, seeking the object specified by a user instruction for segmentation, thereby truly demonstrating the model's ability for active perception and spatial reasoning.

3 The APRS Task and Benchmark

In this section, we formally define the Active Panoramic Referring Segmentation (APRS) task and describe the construction of our benchmark, including the dataset, environment setup, and evaluation protocols.

3.1 Task Formulation

Active Panoramic Referring Segmentation (APRS) requires an agent to search for and segment a target within a continuous 360° environment E based on a language instruction I. Starting from an initial orientation (θ_0, ϕ_0) with a constrained Field of View (FoV), the agent iteratively updates its viewpoint (θ_t, ϕ_t) to explore the environment. At each step t, the agent perceives a visual observation $V_t = \text{Proj}(E, \theta_t, \phi_t, \psi)$ via a perspective projection $\text{Proj}(\cdot)$ with an FoV ψ. Given I, the agent $\mathcal{F}$ maintains a memory state M_t and predicts an action a_t:

$$a_t, M_{t+1} = \mathcal{F}(V_t, I, M_t), \tag{1}$$

where M_{t+1} encapsulates the exploration history updated by the current observation V_t. The action space includes relative movements $a_t = (\Delta\theta_t, \Delta\phi_t)$, denoting the direction of the viewpoint shift, and a termination signal that invokes the prediction of the segmentation mask $\mathcal{S}$ once the target is found.

3.2 Environment and Dataset Construction

Unlike traditional Referring Image Segmentation (RIS) [1, 7], which assumes the target is always within the initial view, the APRS task necessitates active exploration to handle targets that are initially out of sight. To this end, we construct a large-scale benchmark encompassing diverse indoor and outdoor scenes, complemented by complex referring expressions requiring spatial reasoning ability.

Data Sources and Panoramic Simulation. We collect high-resolution panoramas from 360-Indoor [57], PANDORA [58], and SUN360 [59], resulting in 4,971 unique scenes. This scale substantially surpasses existing language-guided 3D benchmarks such as ScanRefer [60] (comprising ∼800 indoor scenes) and offers greater diversity, spanning both indoor and outdoor environments.

To simulate a continuous 360° environment, we employ gnomonic projection to extract local perspective views from equirectangular panoramas (ERPs). Given an orientation (θ, ϕ) and a Field of View (FoV) $\psi = (\psi_h, \psi_v)$, we generate

rectilinear views by back-projecting perspective coordinates onto a unit sphere and subsequently sampling from the ERP. In our implementation, we set the horizontal FoV to $\psi_h = 120°$ and the vertical FoV to $\psi_v = 90°$, aligning with consumer-grade depth cameras. More details are provided in the Appendix.

Human-in-the-Loop Annotation Pipeline. We develop a multi-stage annotation pipeline that combines human expertise with foundation models [61, 62]. First, professional annotators annotate salient targets using bounding boxes. To encourage active exploration, initial camera orientations are directed away from the target, ensuring the target is initially out of the agent's view. Next, based on the spatial relationships between the initial camera orientation, the target, and other anchor objects, annotators are required to construct trajectory-centric spatial referring expressions. These expressions incorporate spatial relations, object attributes, and egocentric cues to guide the agent toward the target. To achieve pixel-level segmentation, we leverage SAM-3 [62] to generate initial masks, followed by manual refinement. Additionally, we employ VLMs (e.g., GPT-5.2 [61]) for semantic cross-checking to ensure the high reliability of our dataset.

Taxonomy of Spatial Referring Expressions. To evaluate diverse spatial reasoning ability, we categorize the spatial referring expressions into four types based on their linguistic and spatial complexity, as shown in Table 1:

- **Egocentric (EGO)**: These instructions specify the target's location relative to the agent's current orientation. The primary challenge lies in mapping linguistic directional cues to precise spatial actions to effectively locate the target. (e.g., *"Turn right to find the sofa."*)
- **Unique-Attribute (UNIQ)**: These instructions identify targets based on distinctive visual attributes (e.g., color or shape) to distinguish them from other objects of the same category within the 360° scene. The core challenge involves fine-grained recognition and grounding of the specified instances across a panoramic view. (e.g., *"The yellow floor lamp in the room."*)
- **Allocentric (ALLO)**: These expressions define the target's position relative to other anchor objects in the 360° environment. The agent must interpret object-to-object spatial relationships to identify the correct target among potential distractors. (e.g., *"The chair opposite the sofa."*)
- **Multi-hop (MULTIHOP)**: These instructions require complex reasoning by integrating both viewpoint adjustments and relative spatial cues. As the most challenging category, they demand sequential decomposition and multi-step reasoning. (e.g., *"Turn around and find the chair next to the table."*)

Supervised Fine-Tuning (SFT) Dataset Construction. Different from the initial APRS dataset annotation Phase, we recruit new annotators to generate exploration trajectories for fine-tuning the agent model. Given an instruction I and an initial camera orientation, annotators are required to find the target by adjusting the FoV based on camera pose parameters (θ, ϕ). We discretize the horizontal $\Delta\theta$ and vertical $\Delta\phi$ action space into multiples of 30°. Horizontal actions include *Small* (30°), *Medium* (60°), *Large* (120°), and *U-turn* (180°); Vertical actions consist of *Small* (30°), *Medium* (60°), and *Large* (90°), facilitating a range from fine alignment to complete FoV shifts. At each step, spatial

Table 1: Taxonomy of Spatial Referring Expressions in the APRS Benchmark. O is the target object; $\{A\}$ denotes visual attributes; R represents spatial relations; v is the initial camera viewpoint; and O_{anc} is the anchor object.

Category	Formal Definition	Description	Number (%)
EGO	$find(O, \{A\}, R, v)$	Relations relative to the initial camera viewpoint	1,365 (18.4%)
UNIQ	$find(O, \{A\})$	Objects with unique visual attributes (e.g., color)	3,213 (43.3%)
ALLO	$find(O, \{A\}, R, O_{anc})$	Relations relative to an external anchor object	2,389 (32.2%)
MULTIHOP	$EGO \circ ALLO$	Multi-step reasoning via viewpoint and spatial cues	453 (6.1%)

coordinates (θ_t, ϕ_t), visual observations V_t, and a memory M_t (detailed later in Sec. 4.1) are recorded. Trajectories exceeding the step limit or deviating significantly from the target are discarded and reassigned to different annotators.

Dataset Statistics. The APRS dataset comprises 7,420 samples across 4,971 scenes, split into training and testing sets at a 7:3 ratio. On average, each scene contains 1.5 instructions spanning four difficulty levels: EGO, UNIQ, ALLO, and MULTIHOP. The average expression length is 7.38 words. These instructions incorporate diverse spatial prepositions (e.g., *"opposite"*, *"behind"*, and *"adjacent to"*), necessitating spatial reasoning and active perception ability to search for and segment targets in the continuous 360° environments.

3.3 Evaluation Metrics

To comprehensively evaluate the performance of the APRS agent, we establish a multi-dimensional evaluation protocol comprising three dimensions: (i) task success, (ii) exploration efficiency, and (iii) segmentation quality.

– **Task Success.** We define the **Success Rate (SR)** to measure the agent's ability to search for and segment the referred target. An episode is considered successful if: (1) the search process terminates at step T; (2) the final viewpoint direction (θ_T, ϕ_T) is within a geodesic distance threshold τ (calculated using the great-circle distance) from the target's ground-truth centroid (θ^*, ϕ^*); and (3) the predicted mask $\mathcal{S}$ achieves an IoU ≥ 0.5. Formally:

$$\mathrm{SR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\mathrm{dist}\{(\theta_{i,T}, \phi_{i,T}), (\theta_i^*, \phi_i^*)\} < \tau \wedge \mathrm{IoU}_i \geq 0.5) \tag{2}$$

where N is the total number of episodes and $\mathbb{I}(\cdot)$ is the indicator function. For each episode, the geodesic distance is defined as:

$$\mathrm{dist}\{(\theta_T, \phi_T), (\theta^*, \phi^*)\} = \arccos\left(\sin\phi_T \sin\phi^* + \cos\phi_T \cos\phi^* \cos(\theta_T - \theta^*)\right) \tag{3}$$

– **Exploration Efficiency.** Efficiency is crucial for active agents. We report the **Average Steps (AS)**, which represents the mean number of action steps taken before termination across successful episodes. Furthermore, we adapt

Success weighted by Path Length (SPL) to the panoramic setting:

$$\text{SPL} = \frac{1}{N}\sum_{i=1}^{N} S_i \cdot \frac{L_i}{\max(L_i, P_i)}, \tag{4}$$

where S_i is the binary success indicator from the SR metric. L_i represents the shortest geodesic distance from the initial orientation (θ_0, ϕ_0) to the target's ground-truth centroid (θ^*, ϕ^*), and P_i is the cumulative great-circle distance traversed by the agent, defined as $P_i = \sum_{t=1}^{T} \text{dist}\{(\theta_t, \phi_t), (\theta_{t-1}, \phi_{t-1})\}$.

– **Segmentation Quality.** Following standard Referring Image Segmentation benchmarks [1, 4, 7], we evaluate the performance using **mean Intersection over Union (mIoU)**. For episodes where SR = 0, the IoU is set to 0.

4 Method

We present **PanoSeeker**, a Vision-Language Agent integrated with a spatial visual memory **EgoSphere**, specifically designed for the APRS task. To ensure non-redundant active search and spatial reasoning, **PanoSeeker** is first trained on an expert-annotated trajectory dataset via Supervised Fine-Tuning (SFT), followed by Reinforcement Learning (RL) to explicitly optimize exploration efficiency. Once the target is found, an **Active Alignment and Segmentation** module is employed for final viewpoint refinement and pixel-level segmentation.

4.1 EgoSphere: Spatial Visual Memory

A primary challenge in APRS is maintaining global spatial consistency under a constrained FoV. Existing memory methods often rely on textual logs or isolated perspective crops—which fail to capture the geometric continuity of a 360° environment and lack trajectory history, leading to redundant exploration (i.e., dead loops). To bridge this gap, we introduce **EgoSphere**, an explicit memory which progressively maps sequential local observations into a geometrically consistent, annotated 360° exploration map to ensure targeted, non-redundant search.

Progressive Canvas Construction. We instantiate the spatial memory as an Equirectangular Panorama (ERP) canvas $M \in \mathbb{R}^{H\times W\times 3}$, initialized as a zero-valued tensor. At each step t, the agent receives a perspective FoV V_t at orientation (θ_t, ϕ_t). To integrate this local observation, we back-project V_t onto a unit sphere and map it to the ERP coordinates:

$$M_t = M_{t-1} \oplus \text{Proj}^{-1}(V_t, \theta_t, \phi_t, \psi), \tag{5}$$

where $\text{Proj}^{-1}(\cdot)$ denotes the inverse gnomonic projection and $\oplus$ signifies the pixel-wise update. This formulation ensures M_t serves as a cumulative visual record, where unexplored regions are explicitly represented as zero-valued frontiers (see Appendix for more gnomonic projection details).

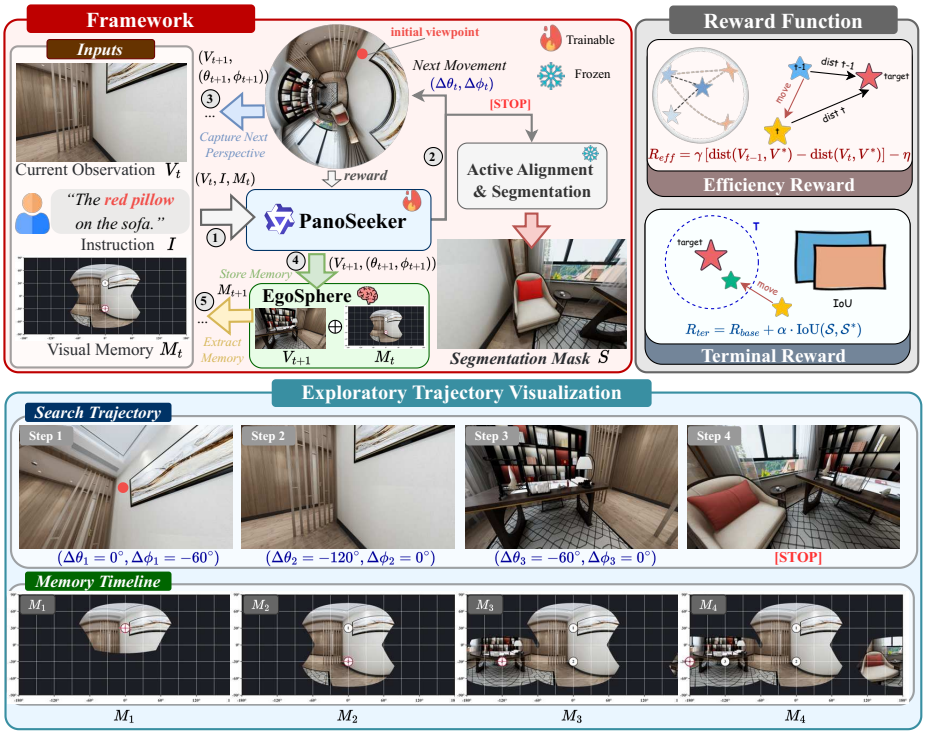

Fig. 2: Overview of the PanoSeeker framework. PanoSeeker integrates a VLM with EgoSphere memory to predict movements $(\Delta\theta_t, \Delta\phi_t)$ from inputs (V_t, I, M_t). Subsequently, an Active Alignment and Segmentation module generates the mask S. Exploration efficiency is guided by efficiency (R_{eff}) and terminal (R_{term}) rewards. Bottom: visualization of a search trajectory and its associated memory accumulation.

Spatial-Aware Visual Prompting. To enable spatial reasoning, we augment M_t with explicit geometric cues that transform the canvas into a structured navigation map. Specifically, we incorporate three types of **visual prompts**: (1) a **central crosshair** that highlights the current FoV center to facilitate egocentric-to-allocentric alignment; (2) a **latitude-longitude grid** at 30° intervals, which serves as a global scale for precise angular estimation; and (3) the **exploration trajectory** $\{(\theta_1, \phi_1), \ldots, (\theta_{t-1}, \phi_{t-1})\}$, rendered as directed paths to track visited regions and prevent redundant scanning. By integrating the recorded memory M_t with the current view V_t, the agent effectively converts the search problem into a visual "fill-in-the-blank" task on the 360° canvas.

4.2 PanoSeeker

The PanoSeeker serves as the spatial reasoning core, leveraging a Vision-Language Model (VLM) [10] to guide camera orientation control. This module processes a

multi-modal input consisting of three key streams: (1) the current local observation V_t; (2) the annotated spatial memory map M_t from EgoSphere; and (3) the instruction I. By jointly attending to the local view (for object detail) and the global map (for path planning), PanoSeeker addresses the search problem by predicting the next optimal action $a_t \in \{(\Delta\theta_t, \Delta\phi_t), \texttt{[STOP]}\}$.

Supervised Fine-Tuning (SFT) Stage. To empower the VLM [10] with active perception ability, we fine-tune PanoSeeker using an expert-annotated trajectory dataset. This stage explicitly trains PanoSeeker to predict the optimal viewpoint adjustment based on the current context. The objective function is defined as:

$$\mathcal{L} = -\mathbb{E}_{\mathcal{T}\sim\mathcal{D}_{\text{expert}}} \sum_{t=1}^{T} \log p_\theta(a_t \mid V_t, M_t, I), \tag{6}$$

where each expert trajectory $\mathcal{T} = \{(V_t, M_t, a_t, I)\}_{t=1}^{T}$ is sampled from the expert dataset $\mathcal{D}_{\text{expert}}$ (detailed in Sec. 3.2). The loss minimizes the negative log-likelihood of the expert action a_t given the learned parameters θ of PanoSeeker, thereby learning the policy for efficient, non-redundant search.

Reinforcement Learning (RL) Stage. To enhance search efficiency, we fine-tune PanoSeeker using Group Relative Policy Optimization (GRPO) [32], which optimizes the policy through relative rewards within sampled groups without a separate critic. To formulate this reinforcement learning process, we first define the reward mechanism to guide the agent's behavior. For a navigation trajectory comprising states $s_t = (V_t, M_t, I)$ of length T, the overall cumulative reward is defined as $r = \sum_{t=1}^{T-1} R_{eff} + R_{term}$. Specifically, the **efficiency reward** $R_{eff} = \gamma[\text{dist}(V_{t-1}, V^*) - \text{dist}(V_t, V^*)] - \eta$ is computed at each step to encourage the agent to follow the shortest geodesic path while penalizing redundant exploration. Upon generating the `[STOP]` token at step T, the **terminal reward** R_{term} assigns a success bonus $R_{base} + \alpha \cdot \text{IoU}$ if the agent successfully finds the target ($\text{dist}(V_T, V^*) < \tau$), and imposes a penalty $-R_{pen}$ otherwise. Given the cumulative reward r evaluated for each trajectory in a sampled group, we maximize the following GRPO objective:

$$\mathbb{E}_{\substack{s_t\sim\mathcal{D},\\ \{a_{t,i}\}_{i=1}^{G}\sim p_{\theta_{\text{old}}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\min\left(\frac{p_\theta(a_{t,i} \mid s_t)}{p_{\theta_{\text{old}}}(a_{t,i} \mid s_t)} A_i, \text{clip}\left(\frac{p_\theta(a_{t,i} \mid s_t)}{p_{\theta_{\text{old}}}(a_{t,i} \mid s_t)}, 1-\epsilon, 1+\epsilon \right) A_i \right) - \beta\, \mathbb{D}_{\text{KL}}(p_\theta \,\|\, p_{\text{ref}}) \right) \right], \tag{7}$$

where p_{ref} is the SFT reference model, G is the group size, and ϵ is the clipping parameter. The advantage A_i for each action $a_{t,i}$ is computed by normalizing its reward r_i within the group: $A_i = (r_i - \text{mean}(\{r_g\}))/\text{std}(\{r_g\})$. β weights the KL-divergence $\mathbb{D}_{\text{KL}}$ for regularization.

4.3 Active Alignment and Segmentation

Upon receiving the `[STOP]` signal, PanoSeeker predicts a bounding box $\mathcal{B}$ by leveraging its inherent VLM [10] grounding ability. To address the challenges of boundary truncation and occlusion—which can lead to unstable mIoU metric evaluation—we implement an active alignment phase. Taking $\mathcal{B}$ as a prompt, we augment SAM-3 [62] to perform concurrent center-seeking, tracking, and segmentation. By applying a final angular adjustment $(\Delta\theta^*, \Delta\phi^*)$, the target's centroid is realigned to the optical center. This ensures the generation of a stable and accurate segmentation mask $\mathcal{S}$ for reliable performance evaluation.

5 Experiments

5.1 Implementation Details

Datasets and Environment. The APRS dataset comprises 7,420 samples across 4,971 scenes, partitioned 7:3 for training and testing. To provide a comprehensive evaluation, we adopt four primary metrics: (1) Success Rate (SR) to measure target localization accuracy; (2) Average Steps (AS) to assess search efficiency; (3) Success weighted by Path Length (SPL) to evaluate trajectory optimality; and (4) mean Intersection over Union (mIoU) to quantify segmentation quality. Each episode is limited to a maximum of $T = 20$ search steps. Each visual observation is rendered with a horizontal Field-of-View (ψ_h) of 120° and a vertical Field-of-View (ψ_v) of 90°, at a resolution of 1024×768 pixels.
Network Architecture: We employ Qwen3-VL-8B-Instruct [10] as the backbone for our PanoSeeker, benefiting from its superior reasoning and grounding ability for the APRS task. Specifically, the model is fine-tuned using a LoRA [63] adapter with a rank $r = 64$, a scaling factor $\alpha = 64$, and a dropout rate of 0.05. Furthermore, the EgoSphere memory is maintained at a resolution of 1024×512, while SAM-3 [62] is used for final segmentation tasks.
Training Details: We implement our model using the DeepSpeed ZeRO-2 framework [64] and conduct training on 4× NVIDIA RTX A6000 GPUs (48GB) with FP16 precision. For Supervised Fine-Tuning (SFT), we employ the AdamW [65] optimizer with a learning rate of 2×10^{-4} and a weight decay of 0.1. A cosine annealing scheduler is applied with a 3% warmup over 3 epochs. The training is performed with a per-GPU batch size of 1 and a gradient accumulation step of 4, resulting in an effective batch size of 16, with a sequence length limit of $4,096$ tokens. In the Reinforcement Learning stage, we leverage the GRPO [32] $(G = 8, \epsilon = 0.2)$ for $6,000$ steps to explicitly optimize exploration efficiency.

5.2 APRS Benchmark Results

In Table 2, we report the quantitative results on our APRS benchmark, comparing PanoSeeker with previous state-of-the-art methods under three search paradigms: (1) *Static Methods*, which directly process the 360° equirectangular panorama (ERP) for referring image segmentation (RIS) in a single forward pass;

Table 2: Quantitative comparison on the APRS benchmark. We compare PanoSeeker with static RIS (processing the entire panorama), heuristic scanning (Upper → Equator → Lower), and active VLM agents. PanoSeeker achieves a significant leap in exploration efficiency (SPL and AS) thanks to its EgoSphere spatial memory and GRPO-optimized [32] policy. The best results are in **bold**.

Method	Venue	Memory	SR (%) ↑	AS ↓	SPL ↑	mIoU (%) ↑
a. Static Methods (Direct Panorama Input)						
VLT [4]	[ICCV'21]	✗	46.7	1.0*	-	34.7
CRIS [3]	[CVPR'22]	✗	55.4	1.0*	-	39.2
LISA [12]	[CVPR'24]	✗	64.1	1.0*	-	44.5
SAM4MLLM [66]	[ECCV'24]	✗	62.3	1.0*	-	46.2
VisionReasoner [15]	[ICLR'26]	✗	66.2	1.0*	-	47.7
b. Heuristic Methods (Pre-defined Scanning, Max 9 steps)†						
VLT [4]	[ICCV'21]	✗	28.4	4.6	0.14	22.1
CRIS [3]	[CVPR'22]	✗	32.1	4.4	0.17	26.4
LISA [12]	[CVPR'24]	✗	42.3	5.2	0.22	34.1
SAM4MLLM [66]	[ECCV'24]	✗	40.5	5.3	0.20	32.5
VisionReasoner [15]	[ICLR'26]	✗	49.5	4.9	0.26	39.9
c. VLM-based Agents (Active Exploration)						
Qwen3-VL-30B-A3B-Thinking [10]	[ArXiv'25]	Text Log	62.9	8.0	0.29	51.0
Gemini-3-Flash [67]	-	Text Log	64.9	6.8	0.31	51.4
GPT-5.2 [61]	-	Text Log	69.1	6.2	0.36	53.2
PanoSeeker (Ours)	-	**EgoSphere**	**75.4**	**4.8**	**0.57**	**55.8**

* *Static methods process the panorama in a single pass (AS=1); SPL is therefore not applicable (−).*
† *Heuristic scanning follows a fixed order: Upper Hemis. → Equator → Lower Hemis.*

(2) *Heuristic Methods*, which follow a predefined scanning paradigm traversing the upper, middle (equatorial), and lower spheres in order, ensuring complete 360° coverage of the space, where the search terminates once the model predicts a segmentation mask; and (3) *VLM-based Agents*, which serve as active controllers using task-specific prompts and textual log memory.

Quantitative Comparison. As shown in Table 2, PanoSeeker achieves the best results across all metrics. Static methods struggle with panoramic image distortion and fail to recognize spatial referring expressions that are hard to identify in a single flat panorama. While heuristic scanning ensures coverage, its fixed path lacks reasoning, leading to many extra steps and a low 0.26 SPL. In contrast, active agents are more suitable for this task because they can reason during exploration. PanoSeeker achieves 75.4% SR and 55.8% mIoU, improving SPL by ∼58% over GPT-5.2 (0.57 vs. 0.36) with only 4.8 steps. These gains are attributed to our EgoSphere's explicit spatial visual memory for path planning and the GRPO-optimized [32] policy that enhances exploration efficiency.

Visualization. Fig. 3 shows PanoSeeker's search trajectories and how the EgoSphere is built step-by-step. Starting from the initial viewpoints (red dots), the agent iteratively adjusts its viewing direction $(\Delta\theta, \Delta\phi)$ to explore the 360° environment for seeking and segmenting targets specified by language instructions. The EgoSphere progressively maps sequential local observations into a 360° map, helping the agent perceive the whole scene and improve exploration efficiency.

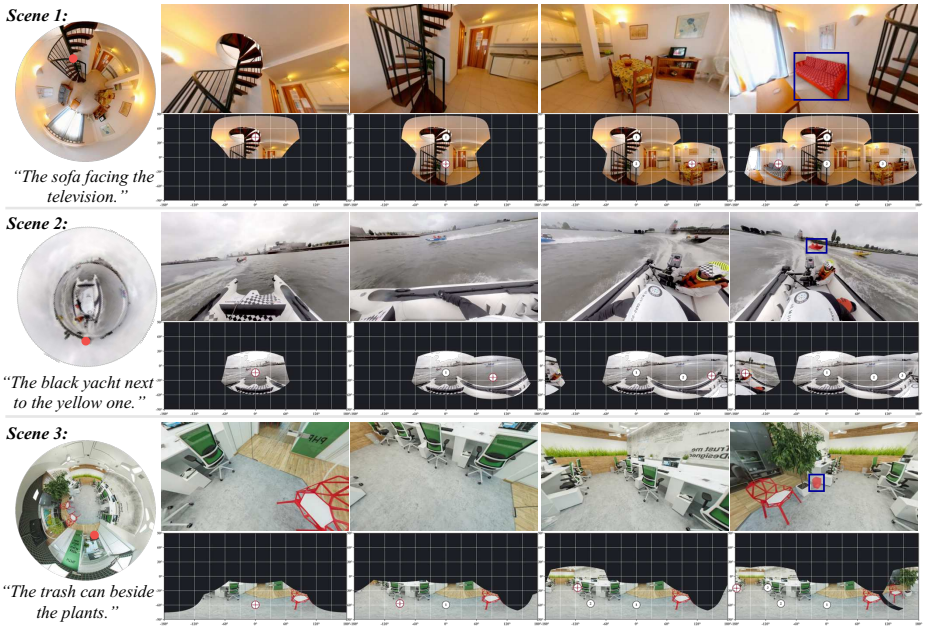

Fig. 3: Visualization Results. Red dots denote initial viewpoints; blue boxes highlight final targets found by PanoSeeker. While these visualizations only show 4-step trajectories, extended scenarios can be found in the Appendix.

These cases are only selected from episodes with the trajectory length of 4; more complex scenarios with longer trajectories are provided in the Appendix.

5.3 Ablation Study

We conduct ablation experiments to verify the effectiveness of all components. All variants use Qwen3-VL-8B [10] as the backbone. All results are processed by the Active Alignment and Segmentation module to ensure stable evaluation.

Table 3: Ablation of **PanoSeeker components**. (a) is a zero-shot baseline using task-specific prompts. We incrementally build upon it to reach our full framework (d).

Variant	Training	EgoSphere	GRPO [32]	SR (%) ↑	AS ↓	SPL ↑	mIoU (%) ↑
(a) Zero-shot	None	✗	✗	41.6	12.4	0.14	28.5
(b) + Memory	None	✓	✗	56.5	8.7	0.26	39.9
(c) + SFT	SFT	✓	✗	70.3	6.2	0.40	50.7
(d) + RL Strategy **(Full)**	SFT+RL	✓	✓	**75.4**	**4.8**	**0.57**	**55.8**

Component Effectiveness. Table 3 shows the contribution of each component to the full PanoSeeker framework. The zero-shot baseline (a) fails to navigate the

Table 4: Comparison of different **memory architectures** in a zero-shot setting. All models are prompted to perform active searching. "Visual Hist." denotes the format of stored historical visual observations. "ERP" denotes Equirectangular Panorama format.

Memory Variant	Text Log	Visual Hist.	SR (%) ↑	AS ↓	SPL ↑	mIoU (%) ↑
(a) Baseline (Zero-shot)	✗	✗	41.6	12.4	0.14	28.5
(b) + Textual Log	✓	✗	50.3	9.9	0.21	35.7
(c) + Visual Buffer	✗	5 frames	46.8	10.5	0.17	32.0
(d) + Hybrid Buffer	✓	5 frames	43.6	11.0	0.15	30.6
(e) + **EgoSphere (Ours)**	✗	**1 ERP canvas**	**56.5**	**8.7**	**0.26**	**39.9**

360° environment effectively, resulting in a high average step (AS=12.4) and low SPL (0.14) due to frequent "dead loops." Integrating EgoSphere (b) significantly improves efficiency, boosting the SPL to 0.26 by providing a persistent visual record that prevents redundant exploration. While the SFT stage (c) further improves the SR to 70.3% by learning experts' search trajectories, the addition of GRPO-based RL (d) achieves the best performance. Specifically, the RL strategy optimizes search trajectories for efficiency, reducing the average steps to 4.8 and reaching the highest SPL of 0.57. This demonstrates that while memory provides the necessary spatial context and historical trajectory, reinforcement learning is key to achieving optimal, non-redundant search paths.

Analysis of Memory Architectures. Table 4 compares various memory strategies in a zero-shot setting. We observe that textual logs (b), visual buffers (c), and hybrid approaches (d) lack sufficient spatial reasoning abilities. Even when past observations are recorded, these sequence-based methods struggle to maintain global context; notably, the hybrid buffer (d) even shows decreased performance due to multimodal information conflicts in the input (43.6 vs. 46.8 SR). In contrast, EgoSphere (e) uses geometric projection to unify local views into a continuous 360° representation, significantly outperforming the next-best variant (Textual Log) by 6.2% in SR and 0.05 in SPL. This highlights that for the APRS task, a geometrically consistent map is far more effective for spatial awareness than simply accumulating temporal logs or isolated observation frames.

6 Conclusion

In this paper, we introduce Active Panoramic Referring Segmentation (APRS), a novel task that requires an agent to actively explore 360° environments for target search and segmentation based on instructions. To address this, we propose PanoSeeker, a memory-augmented agent equipped with EgoSphere—an explicit spatial-visual memory that integrates sequential observations into a unified panoramic representation. By combining Supervised Fine-Tuning with Reinforcement Learning, we explicitly optimize the agent's active perception and spatial reasoning ability. Extensive experiments show that PanoSeeker outperforms existing adapted state-of-the-art baselines, establishing APRS as a scalable, cost-effective benchmark for advancing active perception in Embodied AI.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62472104 and the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600.

References

1. Ding, H., Tang, S., He, S., Liu, C., Wu, Z., Jiang, Y.G.: Multimodal referring segmentation: A survey. arXiv preprint arXiv:2508.00265 (2025)
2. Tang, S., Jie, G., Ding, H., Jiang, Y.G.: ROSE: retrieval-oriented segmentation enhancement. In: CVPR. pp. 7398–7407 (2026)
3. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: CVPR. pp. 11686–11695 (2022)
4. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: ICCV. pp. 16321–16330 (2021)
5. Liu, C., Ding, H., Jiang, X.: GRES: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)
6. Ding, H., Liu, C., He, S., Jiang, X., Jiang, Y.G.: Grex: Generalized referring expression segmentation, comprehension, and generation. IJCV **134**(2), 79 (2026)
7. Ding, H., Liu, C., Wang, S., Jiang, X.: VLT: vision-language transformer and query generation for referring segmentation. IEEE TPAMI **45**(6), 7900–7916 (2022)
8. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
9. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
10. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
11. Shuai, X., Tang, S., Huang, Y., Ding, H., Tao, D.: Psdesigner: Automated graphic design with a human-like creative workflow. In: CVPR. pp. 10165–10175 (2026)
12. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
13. Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G.: Gsva: Generalized segmentation via multimodal large language models. In: CVPR. pp. 3858–3869 (2024)
14. Qian, R., Yin, X., Dou, D.: Reasoning to attend: Try to understand how< seg> token works. In: CVPR. pp. 24722–24731 (2025)
15. Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: Visionreasoner: Unified visual perception and reasoning via reinforcement learning. arXiv e-prints pp. arXiv–2505 (2025)
16. Wang, S., Fang, G., Kong, L., Li, X., Xu, J., Yang, S., Li, Q., Zhu, J., Wang, X.: Pixelthink: Towards efficient chain-of-pixel reasoning. arXiv preprint arXiv:2505.23727 (2025)
17. Wu, W., Chang, T., Li, X., Yin, Q., Hu, Y.: Vision-language navigation: a survey and taxonomy. Neural Computing and Applications **36**(7), 3291–3316 (2024)
18. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d.: Reverie: Remote embodied visual referring expression in real indoor environments. In: CVPR. pp. 9982–9991 (2020)

19. He, S., Ding, H., Jiang, X., Wen, B.: Segpoint: Segment any point cloud via large language model. In: ECCV. pp. 349–367. Springer (2024)
20. Chen, Q., Wu, C., Ji, J., Ma, Y., Yang, D., Sun, X.: Ipdn: Image-enhanced prompt decoding network for 3d referring expression segmentation. In: AAAI. vol. 39, pp. 2132–2140 (2025)
21. He, S., Ding, H.: Refmask3d: Language-guided transformer for 3d referring segmentation. In: ACM MM. pp. 8316–8325 (2024)
22. Qin, Z., Shuai, X., Ding, H.: Scenedesigner: Controllable multi-object image generation with 9-dof pose manipulation. In: NeurIPS (2025)
23. Shuai, X., Ding, H., Qin, Z., Luo, H., Ma, X., Tao, D.: Free-form motion control: Controlling the 6d poses of camera and objects in video generation. In: ICCV. pp. 12449–12458 (2025)
24. Shuai, X., Qin, Z., Ding, H., Tao, D.: Free-form scene editor: Enabling multi-round object manipulation like in a 3d engine. In: AAAI (2025)
25. Li, Z., Luo, H., Shuai, X., Ding, H.: Anyi2v: Animating any conditional image with motion control. In: ICCV. pp. 17302–17311 (2025)
26. He, S., Jie, G., Wang, C., Zhou, Y., Hu, S., Li, G., Ding, H.: ReferSplat: Referring segmentation in 3d gaussian splatting. In: ICML (2025)
27. Liu, Z., Wang, Y., Zheng, S., Pan, T., Liang, L., Fu, Y., Xue, X.: Reasongrounder: Lvlm-guided hierarchical feature splatting for open-vocabulary 3d visual grounding and reasoning. In: CVPR. pp. 3718–3727 (2025)
28. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE TCSVT **35**(7), 6832–6852 (2025)
29. Yan, S., Xu, X., Zhang, R., Hong, L., Chen, W., Zhang, W., Zhang, W.: Panovos: Bridging non-panoramic and panoramic views with transformer for video segmentation. In: ECCV. pp. 346–365. Springer (2024)
30. Zhou, Y., Zhang, T., Zhang, D., Ji, S., Li, X., Qi, L.: Dense360: Dense understanding from omnidirectional panoramas. arXiv preprint arXiv:2506.14471 (2025)
31. Chou, S.H., Chao, W.L., Lai, W.S., Sun, M., Yang, M.H.: Visual question answering on 360deg images. In: WACV. pp. 1607–1616 (2020)
32. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
33. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: ECCV. pp. 108–124. Springer (2016)
34. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: ICCV. pp. 1271–1280 (2017)
35. Li, R., Li, K., Kuo, Y.C., Shu, M., Qi, X., Shen, X., Jia, J.: Referring image segmentation via recurrent refinement networks. In: CVPR. pp. 5745–5753 (2018)
36. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
37. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
38. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)

39. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
40. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
41. Packer, C., Fang, V., Patil, S., Lin, K., Wooders, S., Gonzalez, J.: Memgpt: towards llms as operating systems. (2023)
42. Kang, J., Ji, M., Zhao, Z., Bai, T.: Memory os of ai agent. In: EMNLP. pp. 25972–25981 (2025)
43. Zhong, W., Guo, L., Gao, Q., Ye, H., Wang, Y.: Memorybank: Enhancing large language models with long-term memory. In: AAAI. vol. 38, pp. 19724–19731 (2024)
44. Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., Zhang, Y.: A-mem: Agentic memory for llm agents. arXiv preprint arXiv:2502.12110 (2025)
45. Hu, Z., Zhu, Q., Yan, H., He, Y., Gui, L.: Beyond rag for agent memory: Retrieval by decoupling and aggregation. arXiv preprint arXiv:2602.02007 (2026)
46. Yu, X., Xu, C., Zhang, G., Chen, Z., Zhang, Y., He, Y., Jiang, P.T., Zhang, J., Hu, X., Yan, S.: Vismem: Latent vision memory unlocks potential of vision-language models. arXiv preprint arXiv:2511.11007 (2025)
47. Cirik, V., Berg-Kirkpatrick, T., Morency, L.P.: Refer360: A referring expression recognition dataset in 360 images. In: ACL. pp. 7189–7202 (2020)
48. Ai, H., Cao, Z., Wang, L.: A survey of representation learning, optimization strategies, and applications for omnidirectional vision: H. ai et al. IJCV **133**(8), 4973–5012 (2025)
49. Nakata, A., Yamanaka, T.: 2s-odis: Two-stage omni-directional image synthesis by geometric distortion correction. In: ECCV. pp. 342–356. Springer (2024)
50. Yu, F., Wang, X., Cao, M., Li, G., Shan, Y., Dong, C.: Osrt: Omnidirectional image super-resolution with distortion-aware transformer. In: CVPR. pp. 13283–13292 (2023)
51. Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: ECCV. pp. 707–722 (2018)
52. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360 imagery. NeurIPS **30** (2017)
53. Su, Y.C., Grauman, K.: Kernel transformer networks for compact spherical convolution. In: CVPR. pp. 9442–9451 (2019)
54. Shuai, X., Ding, H., Ma, X., Tu, R., Jiang, Y.G., Tao, D.: A survey of multimodal-guided image editing with text-to-image diffusion models. arXiv preprint arXiv:2406.14555 (2024)
55. Dongfang, Z., Zheng, X., Weng, Z., Lyu, Y., Paudel, D.P., Van Gool, L., Yang, K., Hu, X.: Are multimodal large language models ready for omnidirectional spatial reasoning? arXiv preprint arXiv:2505.11907 (2025)
56. Shuai, X., Li, Z., Ding, H., Tao, D.: Glyphprinter: Region-grouped direct preference optimization for glyph-accurate visual text rendering. In: CVPR. pp. 7674–7683 (2026)
57. Chou, S.H., Sun, C., Chang, W.Y., Hsu, W.T., Sun, M., Fu, J.: 360-indoor: Towards learning real-world objects in 360deg indoor equirectangular images. In: WACV. pp. 845–853 (2020)
58. Xu, H., Zhao, Q., Ma, Y., Li, X., Yuan, P., Feng, B., Yan, C., Dai, F.: Pandora: A panoramic detection dataset for object with orientation. In: ECCV. pp. 237–252. Springer (2022)

59. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: CVPR. pp. 2695–2702. IEEE (2012)
60. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: ECCV. pp. 202–221. Springer (2020)
61. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2026)
62. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
63. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
64. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: SC20: international conference for high performance computing, networking, storage and analysis. pp. 1–16. IEEE (2020)
65. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
66. Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., Chen, C.S.: Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In: ECCV. pp. 323–340. Springer (2024)
67. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)