Panoramic Affordance Prediction

Zixin Zhang[1†], **Chenfei Liao**[1†], **Hongfei Zhang**[1], **Harold H. Chen**[1], **Kanghao Chen**[1], **Zichen Wen**[3], **Litao Guo**[1], **Bin Ren**[4], **Xu Zheng**[1], **Yinchuan Li**[6], **Xuming Hu**[1,2], **Nicu Sebe**[5], **Ying-Cong Chen**[1,2‡]

[1]HKUST(GZ), [2]HKUST, [3]SJTU, [4]MBZUAI, [5]UniTrento, [6]Knowin

[†]Equal contribution, [‡]Corresponding author

Affordance prediction serves as a critical bridge between perception and action in embodied AI. However, existing research is confined to pinhole camera models, which suffer from narrow Fields of View (FoV) and fragmented observations, often missing critical holistic environmental context. In this paper, we present the first exploration into **Panoramic Affordance Prediction**, utilizing 360-degree imagery to capture global spatial relationships and holistic scene understanding. To facilitate this novel task, we first introduce **PAP-12K**, a large-scale benchmark dataset containing over 1,000 ultra-high-resolution (12k, 11904×5952) panoramic images with over 12k carefully annotated QA pairs and affordance masks. Furthermore, we propose **PAP**, a training-free, coarse-to-fine pipeline inspired by the human foveal visual system to tackle the ultra-high resolution and severe distortion inherent in panoramic images. PAP employs recursive visual routing via grid prompting to progressively locate targets, applies an adaptive gaze mechanism to rectify local geometric distortions, and utilizes a cascaded grounding pipeline to extract precise instance-level masks. Experimental results on PAP-12K reveal that existing affordance prediction methods designed for standard perspective images suffer severe performance degradation and fail due to the unique challenges of panoramic vision. In contrast, PAP framework effectively overcomes these obstacles, significantly outperforming state-of-the-art baselines and highlighting the immense potential of panoramic perception for robust embodied intelligence.

Date: March 14, 2026
Project: https://zixinzhang02.github.io/Panoramic-Affordance-Prediction/
Code & Dataset: https://github.com/EnVision-Research/PAP

1 Introduction

Affordance (Gibson, 1977) defines the action possibilities that an environment offers to an agent. As a critical intermediate representation, it serves as the bridge connecting visual perception with actionable capabilities in embodied AI. By explicitly mapping the functionality of the environment—identifying *where* and *how* an agent can interact—affordance prediction provides invaluable semantic information for downstream robotic applications, ranging from task planning and tool usage to complex object manipulation.

To interact seamlessly within complex 3D environments, embodied agents require not just local action cues, but global situational awareness. However, existing research in affordance prediction (Zhang et al., 2025c; Wang et al., 2025a; Wu et al., 2025) is predominantly confined to the "Pinhole Camera Model". This reliance inherently suffers from a restricted Field-of-View (FoV), or "tunnel vision", which fragments the environment into incomplete observations. In practical scenarios, this limitation forces robots to frequently reorient themselves to gather sufficient environmental information, thereby increasing time costs and computational memory burdens. Furthermore, the lack of a holistic view means agents often miss critical environmental cues or potential interaction targets located in their periphery or rear, leading to inefficient task planning and suboptimal decision-making.

Panoramic cameras offer a natural solution to this bottleneck. By capturing a 360° FoV in a single shot, they preserve global spatial relationships and enable holistic scene understanding. Currently, panoramic vision is widely used to help agents "see" the global environment, such as in spatial navigation (Huang and Yeung, 2022; Chen et al., 2024) and scene understanding (Gao et al., 2022; Zheng et al., 2026). However, its potential

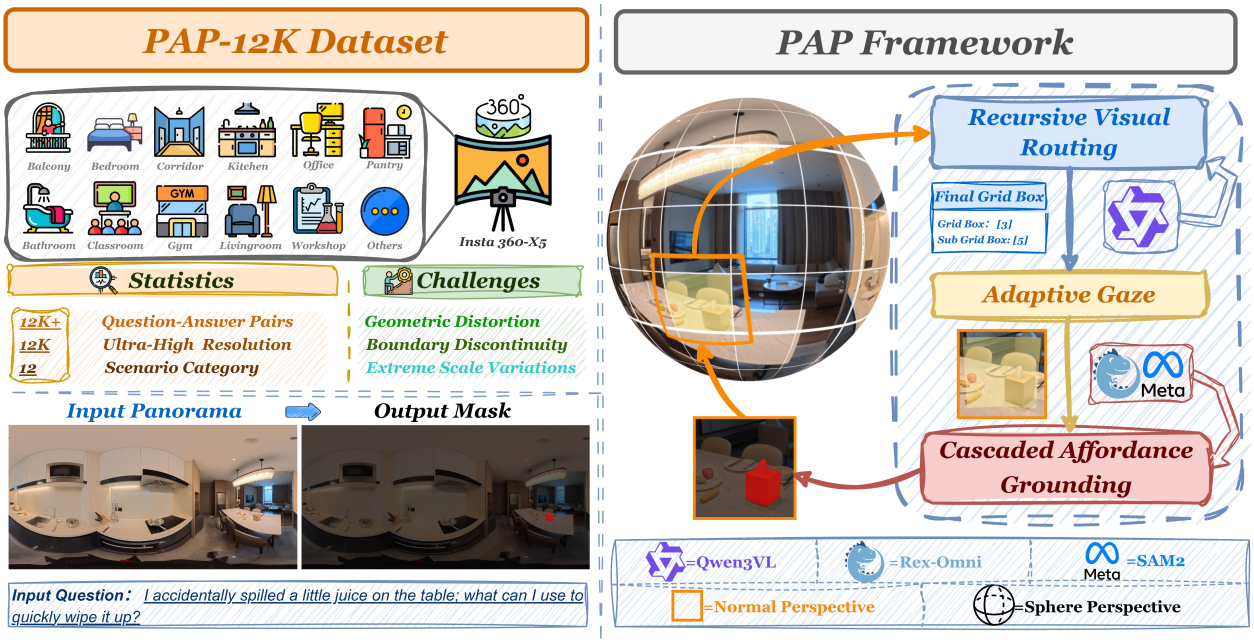


Figure 1 Overview of our work. Left: We introduce PAP-12K, the first large-scale benchmark dedicated to panoramic affordance prediction, featuring ultra-high resolution (12K) imagery, rich reasoning-based QA pairs, and explicitly capturing unique panoramic challenges (geometric distortion, boundary discontinuity, and extreme scale variations). **Right:** We propose the PAP framework, which mimics human foveal vision to tackle these challenges. It employs *Recursive Visual Routing* for efficient coarse localization, an *Adaptive Gaze* mechanism to rectify spatial distortions, and *Cascaded Affordance Grounding* for precise instance-level mask extraction.

to help agents "act" through affordance reasoning is still largely unexplored. Moving affordance prediction into a panoramic view is a critical step forward for embodied AI.

Bridging this gap, we make the ***first exploration*** into the task of ***Panoramic Affordance Prediction***. First, to address the absence of suitable benchmarks in this emerging domain, we introduce **PAP-12K**, a high-resolution, large-scale benchmark dataset designed for panoramic affordance prediction. As shown in Fig. 1, PAP-12K comprises over 1,000 images with ultra-high 12K resolution (11904×5952), captured across hundreds of scenes that span 12 distinct categories, such as daily life, work, and entertainment. PAP-12K are also carefully annotated with over 12k image-question-answer pairs along with corresponding affordance masks, covering diverse objects and tasks. Moreover, we have carefully designed the dataset to incorporate the unique challenges of 360° ERP imagery, including geometric distortion, extreme scale variations, and boundary discontinuity. Extensive experiments on the existing state-of-the-art (SoTA) affordance prediction methods inherently designed for standard perspective images reveal that they suffer severe performance degradation and largely fail when directly applied to panoramic environments. These pipelines struggle to effectively handle the unique challenges posed by 360° imagery, particularly the ultra-high resolution and severe spatial distortions.

To tackle these challenges, we propose **PAP**, a training-free, coarse-to-fine pipeline inspired by the human foveal visual system. PAPmimics how humans visually scan a broad scene before focusing on a specific target. Specifically, the framework operates in three logical stages: (i) To handle extreme scale variations and resolution burdens, we introduce *Recursive Visual Routing via Grid Prompting*. This progressively guides Vision-Language Models (VLMs) to efficiently locate the general area of target tools. (ii) Once the target region is localized, an *Adaptive Gaze* mechanism steps in. It projects this specific spherical region onto a tailored perspective plane, effectively eliminating the geometric distortions and boundary discontinuities inherent to ERP images. (iii) Finally, with a distortion-free local patch secured, a *Cascaded Affordance Grounding* module deploys robust 2D visual foundation models to extract precise, instance-level affordance masks. On our proposed PAP-12K, PAP overcomes the unique challenges of 360° imagery without the need for specialized panoramic fine-tuning, and achieves SoTA performance, underscoring the immense potential of omnidirectional perception for advancing embodied AI. In summary, our main contributions are threefold:

- ❏ **New Task Formulation:** To the best of our knowledge, we make the first exploration into the task of **Panoramic Affordance Prediction**.
- ❏ **Pioneering Benchmark:** We introduce **PAP-12K**, a large-scale, ultra-high-resolution (12K) benchmark dataset explicitly designed for panoramic affordance prediction. It provides rich annotations while encapsulating the unique challenges of 360° ERP imagery.
- ❏ **Novel Framework & SoTA Performance:** We propose **PAP**, a training-free, fovea-inspired pipeline. Our method effectively overcomes challenges of panoramic images, achieving state-of-the-art performance.

2 Related Work

Affordance Prediction: The concept of affordance (Gibson, 1977) defines how embodied agents interact with objects in dynamic physical environments. Predicting affordance is pivotal for downstream embodied tasks, bridging the gap between perception and manipulation by informing task planning (Mo et al., 2021; Wu and Zhao, 2022), object interaction (Hou et al., 2021; Ye et al., 2023), and tool usage (Zhang et al., 2025b; Huang et al., 2024b; Myers et al., 2015). Consequently, it has been widely integrated into numerous robotic systems (Nasiriany et al., 2025; Singh et al., 2025; Xu et al., 2025; Ye et al., 2025). Early learning approaches (Luo et al., 2022; Qian and Fouhey, 2023; Li et al., 2023; Jang et al., 2024) typically formulated the problem as a regression task to predict masks or heatmaps; however, they were often limited to task-specific inferences, lacking robust reasoning capabilities and generalization. With the advent of Vision-Language Models (VLMs) (Hurst et al., 2024; Comanici et al., 2025; Yang et al., 2025), recent methods (Qian et al., 2024; Wu et al., 2025; Wang et al., 2025a) have leveraged fine-tuning strategies to interpret complex task instructions and generate corresponding affordance maps. Most recently, A4-Agent (Zhang et al., 2025c) proposed an agentic framework that decouples reasoning and grounding, achieving strong performance in both areas.

Despite these advancements, existing affordance research relies predominantly on single-frame imagery from pinhole cameras. This restricted "tunnel vision" fragments the environment, forcing robots to perform task planning without a holistic understanding of the scene. Consequently, critical spatial context and potential interaction targets outside the current view are often missed, limiting the applicability of current methodologies in complex real-world scenarios.

Panoramic Vision: Panoramic images offer a 360° field of view, fundamentally distinguishing them from the limited perspective of pinhole cameras. By capturing the entire environment in a single shot, panoramic sensors preserve complete spatial context and global geometric relationships. Despite inherent processing challenges like distortion and scale variation (Lin et al., 2025a), the advantage of "seeing everything at once" has made panoramic vision a crucial enabler for scene understanding (Lin et al., 2025a; Zhou et al., 2025), virtual reality (Lin et al., 2025b), autonomous driving (Qi et al., 2019), and embodied AI (Wan et al., 2025). In Embodied Intelligence, this omnidirectional capability effectively eliminates blind spots. For autonomous navigation and SLAM, it allows agents to maintain stable tracking of landmarks and build consistent maps without frequent reorientation (Chen et al., 2024; Wang et al., 2025b). Beyond navigation, 360° visual inputs significantly enhance an agent's ability in manipulation tasks (Kerr et al., 2025) and facilitate precise localization (Huang et al., 2024a). Furthermore, integrating panoramic vision with multimodal large language models has opened new avenues for spatial reasoning and question answering (Dongfang et al., 2025; Zhang et al., 2025a), achieving a level of environmental completeness that pinhole-based systems cannot match.

Despite the promising applications of panoramic vision in embodied intelligence, its potential for affordance reasoning remains largely unexplored. To bridge this gap, we dive into the field of panoramic affordance prediction. By leveraging the holistic 360° context, we aim to empower embodied agents with a comprehensive understanding of action possibilities across the entire environment, addressing the critical limitations of narrow-field observations.

Table 1 Comparison of our proposed PAP-12K with existing affordance and panoramic datasets. Some Dataset has various resolutions, we adopt the largest resolution for comparison.

Dataset	Affordance Type	# Images	# QAs	Resolution	FoV	Data Source
Affordance Datasets						
UMD (Myers et al., 2015)	Action Category	30k	-	640 × 480	∼ 50°	Real-world Capture
ReasonAff (Wang et al., 2025a)	Instruction & Reasoning	3k	3k	2000 × 1500	∼ 30°	Web Crawled
HANDAL (Guo et al., 2023)	Action Category	200k	-	1920 × 1440	∼ 50°	Real-world Capture
3DOI (Qian and Fouhey, 2023)	Action Category	10k	-	1920 × 1080	∼ 70°	Real-world Capture
RAGNet (Wu et al., 2025)	Instruction & Reasoning	273k	273k	1920 × 1440	∼ 50°	Web Crawled
Panoramic Datasets						
SUN360 (Xiao et al., 2012)	-	67.6k	-	9104 × 4552	360°	Multi-view Stitching
Matterport3D (Chang et al., 2018)	-	10.8k	-	2048 × 1024	360°	Multi-view Stitching
Stanford2D3D (Armeni et al., 2017)	-	1.4k	-	4096 × 2048	360°	Multi-view Stitching
ReplicaPano (Dong et al., 2024)	-	2.7k	-	1024 × 512	360°	Synthetic Rendering
DeepPanoContext (Zhang et al., 2021)	-	1.5k	-	1024 × 512	360°	Synthetic Rendering
Thinking-in-360 (Yu et al., 2025)	-	3k	-	7680 × 3840	360°	Web Crawled
Realsee3D (Li et al., 2025)	-	10k	-	1600 × 800	360°	Multi-view Stitching
Panoramic + Affordance Datasets						
PAP-12K (Ours)	**Instruction & Reasoning**	**1k**	**13k**	**11904 × 5952**	**360°**	**Native 360° Camera**

3 PAP-12K

3.1 Dataset Overview

We introduce PAP-12K, the first and large-scale benchmark dataset dedicated to panoramic affordance prediction. PAP-12K has three key features:

① **Ultra-High Resolution:** To support fine-grained affordance analysis, our images are captured at **an ultra-high resolution of 11904 × 5952**. As shown in Table 1, this resolution not only far exceeds that of existing affordance datasets, but also surpasses standard 360° panoramic datasets. Such a collection of high-quality panoramic images itself provides a valuable resource for the research community. Furthermore, unlike most panoramic datasets relying on multi-view stitching or synthetic rendering from RGB images, PAP-12K is natively captured using professional 360° cameras in real-world environments. This authentic capturing process makes the data much closer to downstream applications, while allowing for the detection of small objects and subtle affordance cues that are often lost in lower-quality imagery.

② **Rich QA Pairs and Masks:** We provide over 13,000 carefully annotated question-answer pairs. Aligning with the latest advancements in affordance datasets (e.g., ReasonAff (Wang et al., 2025a) and RAGNet (Wu et al., 2025)), our QA pairs are specifically designed to require complex reasoning rather than simple perception. Each pair is associated with a precise pixel-level segmentation mask, grounding the affordance in the visual domain. This moves beyond simple classification to require explicit logical deduction and precise localization.

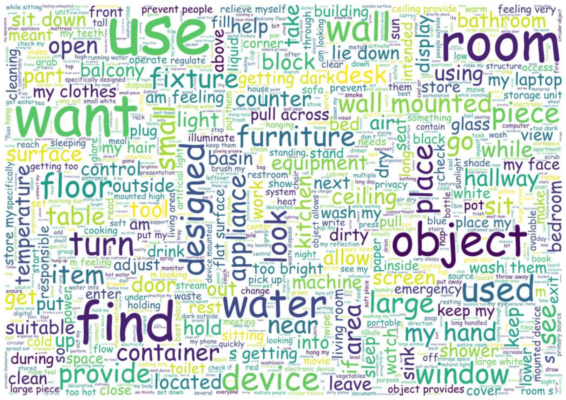


Figure 4 Word Cloud of Questions in PAP-12K.

③ **Panoramic-Specific Challenges:** A key design philosophy of PAP-12K is to explicitly incorporate challenges inherent to 360° panoramic imagery and Equirectangular Projection (ERP). As shown in Fig. 2, PAP-12K includes: *1) Geometric Distortion:* Objects in ERP images suffer from severe stretching, particularly near the poles. PAP-12K includes varied affordance targets with high distortion to evaluate model robustness. *2) Extreme Scale Variations:* Because panoramas capture an unconstrained 360° environment, the visual scale of interactive objects fluctuates drastically. While some objects appear prominent, others occupy an exceptionally small proportion of the total pixels. PAP-12K carefully curates this massive range of scales—with a particular focus on these minute, sub-scale targets—to test a model's ability to localize fine-grained details within a massive global context. *3) Boundary*

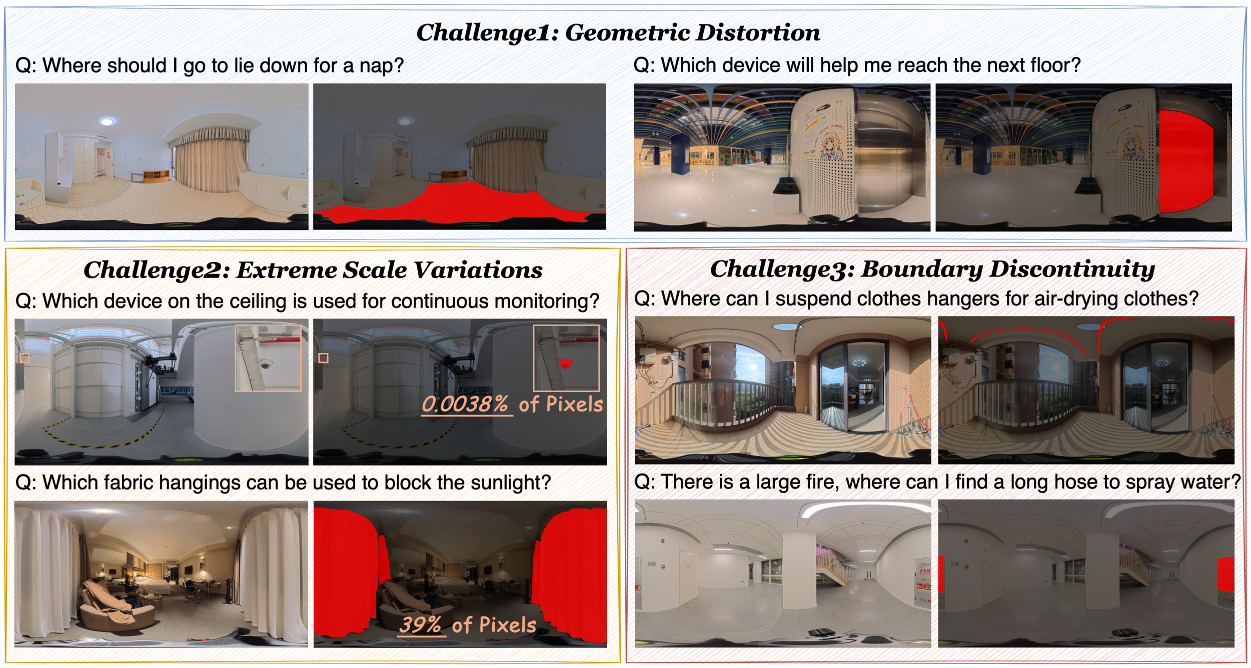

Challenge1: Geometric Distortion
Q: Where should I go to lie down for a nap?
Q: Which device will help me reach the next floor?
Challenge2: Extreme Scale Variations
Q: Which device on the ceiling is used for continuous monitoring?
0.0038% of Pixels
Q: Which fabric hangings can be used to block the sunlight?
39% of Pixels
Challenge3: Boundary Discontinuity
Q: Where can I suspend clothes hangers for air-drying clothes?
Q: There is a large fire, where can I find a long hose to spray water?

Figure 2 PAP-12K specifically features three challenges inherent to 360° panoramic imagery and ERP: (1) Geometric Distortion (e.g., the bed and the elevator); (2) Extreme Scale Variations (e.g., the extremely small security camera and the extremely large curtain); (3) Boundary Discontinuity (e.g., the drying rod and the fire hose).

Discontinuity: In the ERP format, continuous objects or surfaces are often split at the left and right image boundaries. So we have also included some cases where affordance regions wrap around these boundaries, requiring models to reason about continuity rather than treating the image as a static 2D plane.

3.2 Dataset Statistics

As illustrated in Fig. 3, we present the comprehensive statistics of the PAP-12K dataset, categorizing the data across 12 diverse indoor scene types. In total, the dataset comprises 1,003 ultra-high-resolution panoramic images, 6,103 annotated object instances, and 13,493 affordance-related questions. The distribution demonstrates a rich variety of everyday environments. Notably, primary living spaces such as Bedrooms, Bathrooms, Kitchens, and Livingrooms constitute a significant portion of the data across all three hierarchical perspectives (images, instances, and questions). Such a data distribution is intentionally designed to align with the core operational domains of future home assistants and general-purpose robots, bridging the gap between static dataset evaluation and practical downstream robotic applications. This well-distributed and diverse collection therefore provides a robust foundation for training models capable of complex, real-world panoramic affordance reasoning in human-centric environments.

3.3 Dataset Construction

Panoramic Image Collection. All panoramic images were captured using the Insta360-X5, a state-of-the-art professional high-resolution panoramic camera that can capture images at 12K resolution. To ensure high fidelity and simulate realistic robot viewpoints, we employed a professional tripod for stabilization and utilized delayed shooting to eliminate human obstruction from the field of view. Notably, we randomly adjusted the tripod height across different data collection sessions. This variation shifts the camera's equatorial plane, resulting in varying degrees of geometric distortion for objects at different elevations, which poses a realistic challenge for visual perception. For each scene (e.g., the same bedroom), we captured 2 to 4 different images by varying the camera's angle and height, as well as randomly shuffling object arrangements. Consequently, through extensive staging across hundreds of scenes, we collected over 1,000 high-quality, object-rich panoramic images.

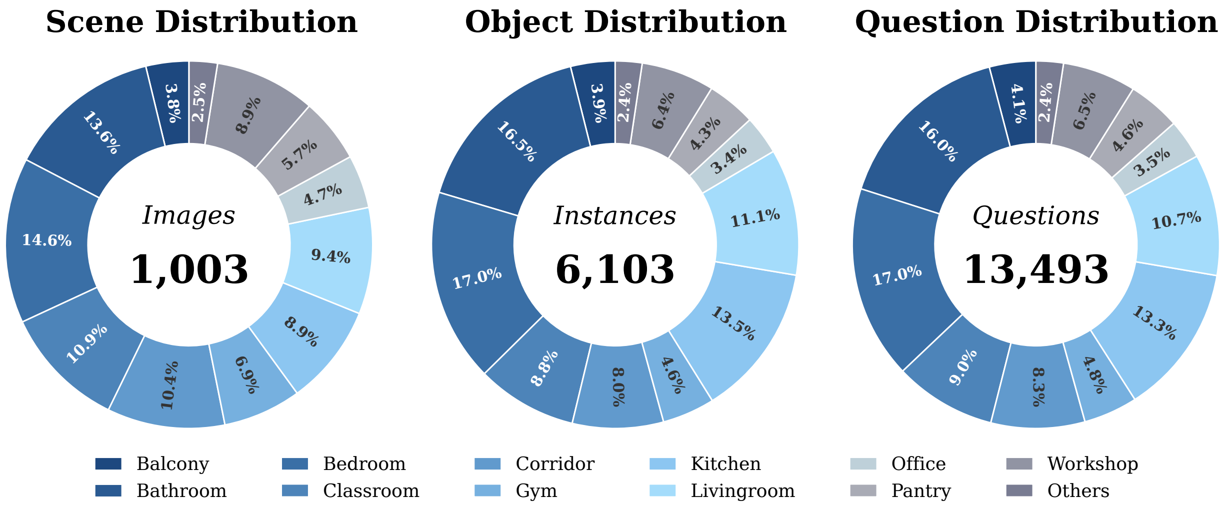


Figure 3 Statistics of the PAP-12K. (Left) Scene distribution with 1,003 high-resolution panoramic images; (Middle) Object distribution featuring 6,103 annotated object instances; (Right) Question distribution comprising 13,493 affordance questions.

Annotation Process. With the high-quality panoramic images collected, the next step involves generating question-answer pairs and corresponding segmentation masks. We decompose this annotation process into two distinct phases. *1) Question Formulation:* The first phase aims to construct valid action/task instruction-object pairs based on the visual content. To streamline this process, we developed an automated agent powered by a Multimodal Large Language Model (MLLM) to generate candidate pairs. While the agent produces a diverse set of instructions, some generated pairs may contain inaccuracies. Therefore, we conducted rigorous multi-round manual verification to ensure high quality and eliminate ambiguity. *2) Mask Segmentation:* The second phase focuses on obtaining precise segmentation masks for the identified objects. This process was done manually with a specifically designed annotation tool. *3) Final Verification:* The final step involves a comprehensive review of the generated question-answer pairs and segmentation masks to ensure high quality and eliminate ambiguity.

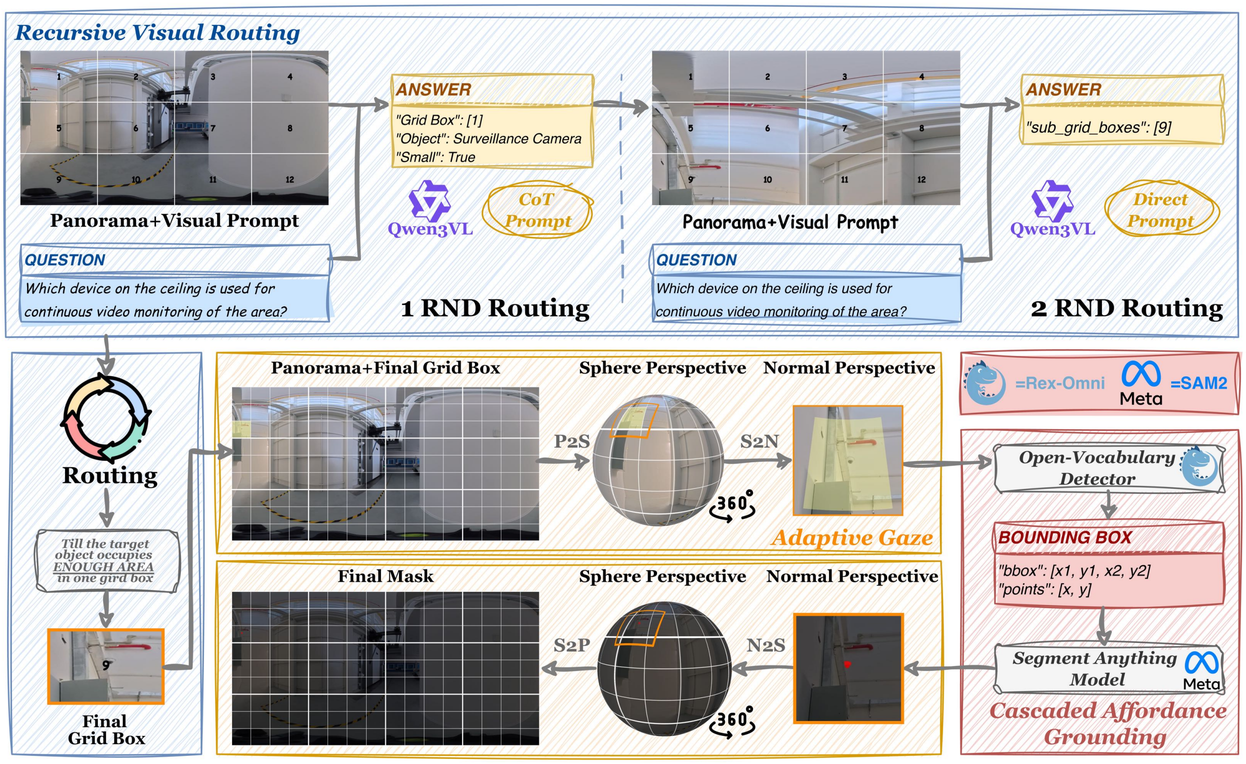
4 PAP

4.1 Overview

Panoramic images offer rich global context but present two critical challenges for affordance prediction: the ultra-high resolution hinders VLMs from precisely capturing fine-grained targets, and Equirectangular Projection (ERP) distortions create a severe domain gap that degrades foundation model performance. To tackle these, we propose **PAP**, a training-free, coarse-to-fine pipeline inspired by the human foveal visual system. Instead of processing 360° scenes with uniform acuity, humans use peripheral vision to locate regions of interest, direct their gaze for a clear view, and perform detailed parsing. Mirroring this, PAP operates in three stages (Fig. 5): *Recursive Visual Routing via Grid Prompting* (Sec. 4.2) leverages VLMs for coarse spatial localization. Then, *Adaptive Gaze* (Sec. 4.3) eliminates geometric distortions to bridge the domain gap. Finally, a *Cascaded Affordance Grounding* pipeline (Sec. 4.4) deploys an Open-Vocabulary Detector (OVD) and the Segment Anything Model (SAM) within the rectified patch for precise instance-level segmentation.

4.2 Recursive Visual Routing via Grid Prompting

While existing VLMs possess strong semantic reasoning capabilities for deducing implicit affordances (i.e., inferring *what* tool is needed for a task), they inherently struggle with precise spatial grounding, especially in high-resolution panoramic images due to token length constraints and attention dilution. To compensate for this grounding deficiency and avoid forcing the VLM to regress explicit continuous coordinates, we introduce



"Grid Box": [1]
"Object": Surveillance Camera
"Small": True

"sub_grid_boxes": [9]

Which device on the ceiling is used for continuous video monitoring of the area?

Which device on the ceiling is used for continuous video monitoring of the area?

"bbox": [x1, y1, x2, y2]
"points": [x, y]

Figure 5 Illustration of the PAP framework.

a **Visual Grid Prompting** mechanism. Given an input ERP image $\mathcal{I}_{ERP}$ and a task description $\mathcal{T}$, we overlay a 4×3 numerical grid (indexed 1 to 12) onto the image. The VLM is then tasked to output a concise textual object description $\mathcal{T}_{obj}$ that can be used to do the task $\mathcal{T}$, and select the specific grid index containing the target.This visual prompt effectively transforms the complex continuous localization task into a discrete, multi-modal multiple-choice question, successfully bridging the gap between semantic reasoning and spatial grounding.

However, a single-pass grid selection is often insufficient for tiny objects within a massive 360° view. To address the extreme scale variations of objects, we further propose **Recursive Visual Routing** that guides the VLM to dynamically "zoom in" based on the target's relative scale: ❶ Scale-Salient Targets: If the target spans multiple grids, the current resolution is sufficient, and the recursion terminates. ❷ *Sub-Scale Targets:* If confined to a single grid, the object is too small. We crop this grid, overlay a new grid, and infer recursively. Crucially, to ensure computational efficiency, input images are aggressively downsampled during this recursion. Initially, the global panorama is severely downsampled to extract broad contextual cues. After this first step, the system recursively zooms into smaller cropped grids and reduces the downsampling ratio. This guided localization efficiently filters out irrelevant 360° backgrounds, securing a reliable coarse bounding region for subsequent modules without the massive overhead of exhaustive high-resolution processing.

4.3 Adaptive Gaze

Once the local grid containing the target is coarsely isolated, directly giving precise segmentation masks is still challenging. This is due to the pronounced challenges inherent in the ERP format (as discussed in Sec. 3.1), which creates a significant domain gap between panoramic images and standard perspective images. To bridge this gap, we introduce the **Adaptive Gaze**, simulating the human action of turning one's head to focus directly on a target. This mechanism adaptively adjusts both its viewing direction and scope to match the routed grid. Specifically, we first direct the *"gaze"* by aligning the camera's principal point (point of tangency) with the latitude and longitude (ϕ_c, θ_c) of the grid's center. Then, we adjust the *"focus"* by adaptively scaling the Field of View (FoV) based on the grid's location and dimensions to perfectly enclose the target region. By projecting this local spherical region onto a tangent plane, we yield a perspective image

Table 2 Comparison on PAP-12K. **Best** results and <u>Second-best</u> results are highlighted in bold and underline.

Method	gIoU↑	cIoU↑	P_{50} ↑	P_{50-95} ↑	Inference Time
OV-Seg (Liang et al., 2023)	29.48	17.85	32.00	18.80	∼8s
LISA (Lai et al., 2024)	15.21	16.34	13.66	8.30	∼7s
VisionReasoner (Liu et al., 2025)	49.33	44.64	51.06	38.06	∼12s
AffordanceVLM (Wu et al., 2025)	9.66	13.11	8.96	5.41	∼7.8s
Affordance-R1 (Wang et al., 2025a)	51.80	<u>50.32</u>	55.47	40.70	∼10.4s
A4-Agent (Zhang et al., 2025c)	<u>62.55</u>	49.97	<u>67.09</u>	<u>54.28</u>	∼11.8s
PAP (Ours)	**71.56**	**62.30**	**75.49**	**64.97**	∼10s

$\mathcal{I}_{persp}$ tailored specifically to the target.

This spherical-to-perspective projection acts as a **training-free domain adapter** that elegantly resolves key panoramic challenges: 1) It eliminates *geometric distortion* by mapping the local curved surface to a flat tangent plane. 2) The adaptive FoV naturally mitigates *extreme scale variations* by magnifying sub-scale targets to standard resolutions and maintaining fine-grained details. 3) It seamlessly handles *boundary discontinuity* as the projection operates intrinsically on the continuous spherical manifold; severed targets on the 2D ERP images are reconstructed as whole objects in the perspective crop. By overcoming these inherent ERP limitations, we ensure seamless integration with powerful 2D visual foundation models without the need for panoramic-specific fine-tuning.

4.4 Cascaded Affordance Grounding

With the reasoning and coarse localization handled by the Recursive Visual Routing, and the geometric distortion eliminated by the Adaptive Gaze, the pipeline proceeds to pixel-level grounding. Following recent success in decoupling 2D affordance grounding into distinct reasoning and perception modules (Zhang et al., 2025c), we deploy a cascaded pipeline on the high-resolution local perspective image $\mathcal{I}_{persp}$.

We first deploy an Open-Vocabulary Detector (OVD) to parse the deduced explicit object description $\mathcal{T}_{obj}$ (generated by the Recursive Visual Routing) and scan the refined perspective image $\mathcal{I}_{persp}$ to generate a bounding box $\mathcal{B}$ and key points $\mathcal{P}$. By utilizing the specific object name rather than the complex, implicit task description $\mathcal{T}$, and because the Recursive Visual Routing has already filtered out the vast majority of the 360° background, the OVD is relieved of the immense burden of semantic guessing and searching the entire complex scene. This results in significantly higher detection accuracy and fewer false positives. Subsequently, this bounding box $\mathcal{B}$ and key points $\mathcal{P}$ serve as a dense spatial prompt for the SAM. Leveraging SAM's robust zero-shot segmentation capabilities, we extract a highly accurate, instance-level segmentation mask $\mathcal{M}_{persp}$ along the target's boundaries. Finally, to fulfill the end-to-end panoramic affordance prediction task, the predicted perspective mask $\mathcal{M}_{persp}$ is re-mapped back to the original ERP space via an inverse perspective-to-spherical projection transformation, yielding the final panoramic mask $\mathcal{M}_{ERP}$. This synergistic design naturally breaks down the formidable 360° grounding problem into manageable stages, playing precisely to the strengths of each foundation model.

5 Experiments

5.1 Experimental Settings

Baseline Methods: To the best of our knowledge, our method is the first framework dedicated to panoramic affordance prediction. Due to the lack of direct panoramic baselines, we validate the effectiveness of our approach by benchmarking it against state-of-the-art methods designed for perspective imagery, including A4-Agent (Zhang et al., 2025c), Affordance-R1 (Wang et al., 2025a), and AffordanceVLM (Wu et al., 2025).

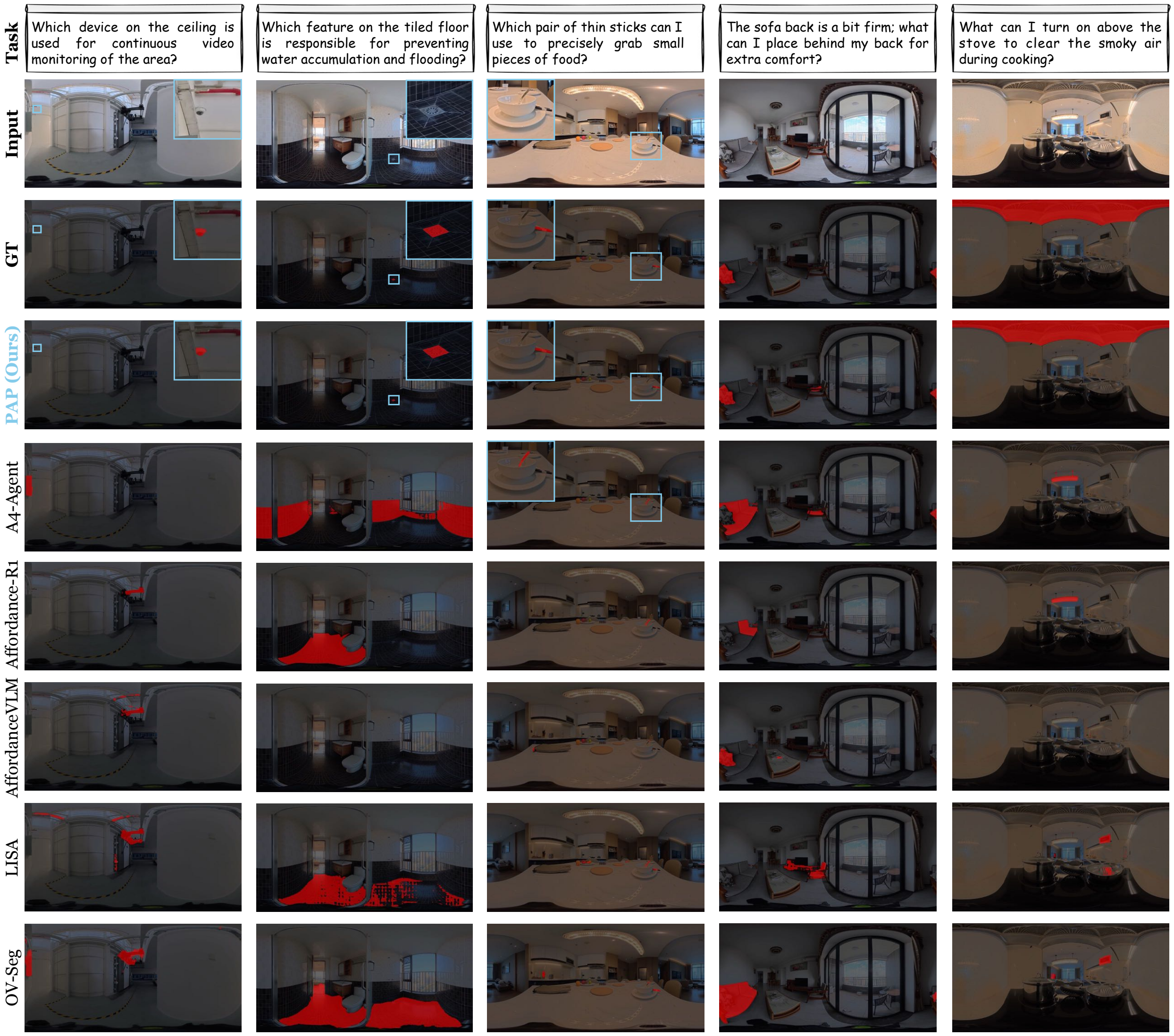

Task
Which device on the ceiling is used for continuous video monitoring of the area?
Which feature on the tiled floor is responsible for preventing water accumulation and flooding?
Which pair of thin sticks can I use to precisely grab small pieces of food?
The sofa back is a bit firm; what can I place behind my back for extra comfort?
What can I turn on above the stove to clear the smoky air during cooking?
Input
GT
PAP (Ours)
A4-Agent
Affordance-R1
AffordanceVLM
LISA
OV-Seg

Figure 6 Qualitative comparison on PAP-12K. We highlight the predicted affordance regions in red. The extremely small objects are enlarged for better visualization.

Furthermore, following the evaluation protocols in (Zhang et al., 2025c; Wang et al., 2025a), we extend our comparison to include general Open Vocabulary Segmentation models, such as Vision Reasoner (Liu et al., 2025),OVSeg (Liang et al., 2023), and LISA (Lai et al., 2024).

Evaluation Metrics: Following the standard evaluation protocol in affordance prediction (Wang et al., 2025a; Zhang et al., 2025c), we adopt four complementary metrics to comprehensively assess prediction quality, including: *1) gIoU (Generalized IoU):* The average Intersection-over-Union across all test samples, measuring the overall segmentation quality of the predicted affordance regions. *2) cIoU (Cumulative IoU):* The cumulative intersection over cumulative union across the entire dataset, providing a dataset-level quality measure that is less sensitive to the size of individual objects. *3) P@50 (Precision at IoU=0.5):* The percentage of predictions with an IoU score exceeding 0.5, evaluating the model's ability to generate high-quality predictions. *4) P@50:95:* The average precision calculated across a range of IoU thresholds from 0.5 to 0.95 with 0.05 increments, providing a stricter and more fine-grained assessment of segmentation accuracy.

Implementation Details: For our PAP framework, we adopt Qwen3-VL-32B (Bai et al., 2025) as the backbone of our VLM, and following (Zhang et al., 2025c), we adopt Rex-Omni (Jiang et al., 2025) as the OVD model and the SAM-2-Large (Ravi et al., 2024) as the Segmentation model. During the Recursive

Table 3 Comparison on different difficulty levels of PAP-12K. **Best** results are highlighted in bold. <u>Second-best</u> results are highlighted in underline.

Method	PAP-12K-Hard				PAP-12K-Normal			
	gIoU↑	cIoU↑	P_{50}↑	P_{50-95}↑	gIoU↑	cIoU↑	P_{50}↑	P_{50-95}↑
OV-Seg (Liang et al., 2023)	10.66	9.18	10.32	4.90	37.66	21.72	41.44	24.84
LISA (Lai et al., 2024)	4.75	13.04	3.23	2.01	19.75	17.24	18.20	11.03
VisionReasoner (Liu et al., 2025)	25.96	29.74	24.91	15.00	59.50	49.59	62.44	48.10
AffordanceVLM (Wu et al., 2025)	4.94	15.94	3.59	1.92	11.72	12.36	11.30	6.93
Affordance-R1 (Wang et al., 2025a)	33.31	<u>40.84</u>	36.26	21.39	59.85	52.97	63.82	49.11
A4-Agent (Zhang et al., 2025c)	<u>42.75</u>	36.42	<u>46.48</u>	<u>30.38</u>	<u>70.91</u>	<u>54.94</u>	<u>75.79</u>	<u>64.36</u>
PAP (Ours)	**60.35**	**52.59**	**63.82**	**52.17**	**76.40**	**65.20**	**80.52**	**70.49**

Visual Routing stage, we apply the dynamic resolution adaptation strategy to balance computational cost and visual details. Specifically, in the first routing round, the full panorama is downsampled to 2000×1000 (a $\sim 1/6$ scaling ratio) before being fed into the VLM. If the target is sub-scale and requires a second routing round, the cropped grid region is downsampled to 1500×1000 (a $\sim 1/2$ scaling ratio). This design elegantly balances computational efficiency and visual performance, ensuring that as the routing progresses into smaller regions, the VLM receives images with progressively higher relative pixel fidelity without exceeding strict token constraints.

5.2 Main Results

Overall Performance. Table 2 presents the overall quantitative comparison between our proposed PAP and state-of-the-art baselines. PAP significantly outperforms all existing methods across all evaluation metrics by a large margin. Specifically, PAP achieves a gIoU of 71.56% and a cIoU of 62.30%, surpassing the second-best method, A4-Agent, by absolute margins of 9.01% and 12.33%, respectively. In terms of precision, our method obtains 75.49% on P_{50} and 64.97% on P_{50-95}, demonstrating its superior capability in precise affordance localization and segmentation compared to recent strong baselines. The substantial improvements indicate that our approach effectively captures complex panoramic affordance relationships. Fig. 6 presents several representative cases. As illustrated, our method robustly accomplishes the task even in challenging scenarios—such as when objects are excessively large or small, split across boundaries, or severely distorted—whereas other methods struggle.

Performance on Different Difficulty Levels. Furthermore, we quantitatively evaluate the performance of different models under varying difficulty levels. We partition PAP-12K into *Hard* and *Normal* subsets based on the following criteria: 1) The object is exceptionally large or small, with the mask occupying more than 30% or less than 0.1% of the entire image. 2) The object is truncated across the left and right boundaries, resulting in the mask appearing on both sides of the image. This process identifies approximately 30% of the cases as the *Hard* subset, leaving the remainder as the *Normal* subset. Table 3 illustrates the comparison across these difficulty levels. Notably, in the *Hard* subset, our method exhibits an even more pronounced advantage, outperforming the second-best method (A4-Agent) by absolute margins of 17.60% and 16.17% in terms of gIoU and cIoU. In terms of precision, PAP leads A4-Agent by absolute margins of 17.34% and 21.79% on P_{50} and P_{50-95}. This demonstrates that our method effectively addresses the inherent challenges of panoramic ERP images such as extreme scale variation and boundary discontinuity.

5.3 Ablation Studies

To validate the effectiveness of our approach, we conducted extensive ablation studies. For computational efficiency, we randomly sampled 10% of the PAP-12K for these experiments. We evaluate three key components: Prompt Style, Recursive Visual Routing, and Adaptive Gaze. In this section, we analyze the overall impact of integrating versus omitting each module. Detailed hyperparameter analyses and additional analytical experiments are deferred to the Appendix (Sec. A).

Table 4 Ablation on Prompt Style.

CoT	VGP	gIoU↑	cIoU↑	$\mathbf{P_{50}}$ ↑	$\mathbf{P_{50-95}}$ ↑
✗	✗	57.70	57.68	60.76	50.52
✗	✓	69.56	61.05	74.02	61.90
✓	✗	67.22	58.75	70.61	60.51
✓	✓	72.69	63.85	76.29	66.13

Table 5 Ablation on Adaptive Gaze.

Method	gIoU↑	cIoU↑	$\mathbf{P_{50}}$ ↑	$\mathbf{P_{50-95}}$ ↑
w/o AG	64.99	55.43	68.43	56.37
w/ AG	72.69	63.85	76.29	66.13
Δ	7.70	8.42	7.77	9.76

Table 6 Ablation on RVR.

Subset	Method	gIoU↑	cIoU↑	$\mathbf{P_{50}}$ ↑	$\mathbf{P_{50-95}}$ ↑
All	w/o RVR	70.56	57.05	74.43	63.23
	w/ RVR	72.69	63.85	76.29	66.13
	Δ	2.13	6.80	1.86	2.90
Hard	w/o RVR	57.37	34.81	61.00	47.65
	w/ RVR	62.40	47.11	64.89	53.64
	Δ	5.03	12.30	3.89	5.99
Normal	w/o RVR	76.29	63.55	80.26	69.99
	w/ RVR	77.07	68.05	81.15	71.45
	Δ	0.78	4.50	0.89	1.46

Impact of Prompt Style. We investigate the impact of different prompting strategies, specifically *Chain-of-Thought (CoT)* and the proposed *Visual Grid Prompting (VGP)*. As shown in Table 4, incorporating CoT reasoning steadily improves the model's performance, demonstrating that step-by-step logical deduction aids in complex spatial prediction. More importantly, the explicit visual grid overlay proves to be an even more crucial component. When replacing the visual grid with a purely textual spatial description (e.g., verbally describing the image divisions to the VLM). Comparing Row 1 vs. Row 2, and Row 3 vs. Row 4, we observe a drastic performance drop across all metrics. This confirms that while VLMs possess strong semantic reasoning capabilities, they struggle to map abstract textual spatial divisions to complex visual features in ultra-high-resolution panoramas without explicit visual anchors. Our Visual Grid Prompting successfully bridges this gap by providing concrete reference points, effectively grounding the VLM's reasoning process into a manageable multi-modal discrete choice task.

Impact of Recursive Visual Routing. To evaluate the efficacy of our Recursive Visual Routing (RVR) module, we compare the full model against a baseline that relies solely on a single-step grid routing across different difficulty subsets. As shown in Table 6, integrating RVR yields consistent improvements on the entire dataset. More importantly, the performance gains are significantly more pronounced on the "Hard" subset, where cIoU and gIoU surge by 12.30% and 5.03%, respectively, compared to the modest increases of 4.50% and 0.78% on the "Normal" subset. This disparity demonstrates that due to the extreme object scale variations inherent in 360° scenes, a single-step routing approach tends to produce excessively large localization regions and struggles to precisely locate challenging targets. Conversely, our RVR module overcomes this limitation by iteratively zooming in on the target, ensuring robust coarse localization even for the most difficult objects regardless of their scales.

Impact of Adaptive Gaze. A core insight of our pipeline is the training-free domain adaptation via Adaptive Gaze (AG). We ablate this module by directly applying the OVD and SAM on the cropped ERP regions without AG. The results show a significant decline in accuracy. This validates that the inherent spatial distortions of ERP formats introduce a severe domain shift that disrupts the pre-trained priors of 2D foundation models. Our AG successfully eliminates this distortion, seamlessly aligning the data domains.

6 Conclusion

In this paper, we present the first exploration into **Panoramic Affordance Prediction**, bridging the critical gap between holistic 360° scene understanding and actionable intelligence in embodied AI. We first introduce PAP-12K, a pioneering, large-scale benchmark featuring over 1,000 natively captured ultra-high-resolution (12K) panoramic images and more than 12K reasoning-based QA pairs with precise affordance masks. Furthermore, to address the profound challenges of panoramic vision, we propose PAP, a training-free framework inspired by the human foveal visual system. Extensive experiments demonstrate that PAP significantly outperforms existing state-of-the-art methods, particularly in highly challenging scenarios. We hope our dataset and framework will serve as a foundational stepping stone, inspiring future research toward the combination of panoramic vision and embodied intelligence.

References

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *7th IEEE International Conference on 3D Vision, 3DV 2017*, pages 667–676. Institute of Electrical and Electronics Engineers Inc., 2018.

Yichen Chen, Yuqi Pan, Ruyu Liu, Haoyu Zhang, Guodao Zhang, Bo Sun, and Jianhua Zhang. 360orb-slam: A visual slam system for panoramic images with depth completion network. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 717–722. IEEE, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yuan Dong, Chuan Fang, Liefeng Bo, Zilong Dong, and Ping Tan. Panocontext-former: Panoramic total scene understanding with a transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28087–28097, 2024.

Zihao Dongfang, Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Danda Pani Paudel, Luc Van Gool, Kailun Yang, and Xuming Hu. Are multimodal large language models ready for omnidirectional spatial reasoning? *arXiv preprint arXiv:2505.11907*, 2025.

Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. Review on panoramic imaging and its applications in scene understanding. *IEEE Transactions on Instrumentation and Measurement*, 71:1–34, 2022.

James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11428–11435. IEEE, 2023.

Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 495–504, 2021.

Huajian Huang and Sai-Kit Yeung. 360vo: Visual odometry using a single 360 camera. In *2022 international conference on robotics and automation (icra)*, pages 5594–5600. IEEE, 2022.

Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, and Sai-Kit Yeung. 360loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22314–22324, 2024a.

Siyuan Huang, Iaroslav Ponomarenko, Zhengkai Jiang, Xiaoqi Li, Xiaobin Hu, Peng Gao, Hongsheng Li, and Hao Dong. Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7580–7587. IEEE, 2024b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ji Ha Jang, Hoigi Seo, and Se Young Chun. Intra: Interaction relationship-aware weakly supervised affordance grounding. In *European Conference on Computer Vision*, pages 18–34. Springer, 2024.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction, 2025. URL https://arxiv.org/abs/2510.12798.

Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a bc-rl perception-action loop. *arXiv preprint arXiv:2506.10968*, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024.

Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023.

Linyuan Li, Yan Wu, Xi Li, Lingli Wang, Tong Rao, Jie Zhou, Cihui Pan, and Xinchen Hui. Realsee3d: A large-scale multi-view rgb-d dataset of indoor scenes (version 1.0), 2025. URL https://doi.org/10.5281/zenodo.17826243.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023.

Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Xiangtai Li, Wenjie Jiang, Bo Du, Dacheng Tao, Ming-Hsuan Yang, and Lu Qi. One flight over the gap: A survey from perspective to panoramic vision. *arXiv preprint*, 2025a.

Xin Lin, Shi Luo, Xiaojun Shan, Xiaoyu Zhou, Chao Ren, Lu Qi, Ming-Hsuan Yang, and Nuno Vasconcelos. Hqgs: High-quality novel view synthesis with gaussian splatting in degraded scenes. In *ICLR*, 2025b.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1374–1381. IEEE, 2015.

Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8249–8257. IEEE, 2025.

Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *CVPR*, 2019.

Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21753–21763, 2023.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Anukriti Singh, Kasra Torshizi, Khuzema Habib, Kelin Yu, Ruohan Gao, and Pratap Tokekar. Afford2act: Affordance-guided automatic keypoint selection for generalizable and lightweight robotic manipulation. *arXiv preprint arXiv:2510.01433*, 2025.

Zhaoliang Wan, Zetong Bi, Zida Zhou, Hao Ren, Yiming Zeng, Yihan Li, Lu Qi, Xu Yang, Ming-Hsuan Yang, and Hui Cheng. Rapid hand: A robust, affordable, perception-integrated, dexterous manipulation platform for generalist robot autonomy. *arXiv preprint arXiv:2506.07490*, 2025.

Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. *arXiv preprint arXiv:2508.06206*, 2025a.

Sen Wang, Dongliang Zhou, Liang Xie, Chao Xu, Ye Yan, and Erwei Yin. Panogen++: Domain-adapted text-guided panoramic environment generation for vision-and-language navigation. *Neural Networks*, 187:107320, 2025b.

Dongming Wu, Yanping Fu, Saike Huang, Yingfei Liu, Fan Jia, Nian Liu, Feng Dai, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, et al. Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11980–11990, 2025.

Ruihai Wu and Yan Zhao. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. In *International Conference on Learning Representations (ICLR), 2022*, 2022.

Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2695–2702. IEEE, 2012.

Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, et al. A0: An affordance-aware hierarchical model for general robotic manipulation. *arXiv preprint arXiv:2504.12636*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Angen Ye, Zeyu Zhang, Boyuan Wang, Xiaofeng Wang, Dapeng Zhang, and Zheng Zhu. Vla-r1: Enhancing reasoning in vision-language-action models. *arXiv preprint arXiv:2510.01623*, 2025.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023.

Heyang Yu, Yinan Han, Xiangyu Zhang, Baiqiao Yin, Bowen Chang, Xiangyu Han, Xinhao Liu, Jing Zhang, Marco Pavone, Chen Feng, et al. Thinking in 360 {\deg}: Humanoid visual search in the wild. *arXiv preprint arXiv:2511.20351*, 2025.

Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2021.

Xinshen Zhang, Zhen Ye, and Xu Zheng. Towards omnidirectional reasoning with 360-r1: A dataset, benchmark, and grpo-based method. *arXiv preprint arXiv:2505.14197*, 2025a.

Zixin Zhang, Kanghao Chen, Xingwang Lin, Lutao Jiang, Xu Zheng, Yuanhuiyi Lyu, Litao Guo, Yinchuan Li, and Ying-Cong Chen. Phystoolbench: Benchmarking physical tool understanding for mllms. *arXiv preprint arXiv:2510.09507*, 2025b.

Zixin Zhang, Kanghao Chen, Hanqing Wang, Hongfei Zhang, Harold Haodong Chen, Chenfei Liao, Litao Guo, and Ying-Cong Chen. A4-agent: An agentic framework for zero-shot affordance reasoning. *arXiv preprint arXiv:2512.14442*, 2025c.

Xu Zheng, Chenfei Liao, Ziqiao Weng, Kaiyu Lei, Zihao Dongfang, Haocong He, Yuanhuiyi Lyu, Lutao Jiang, Lu Qi, Li Chen, Danda Pani Paudel, Kailun Yang, Linfeng Zhang, Luc Van Gool, and Xuming Hu. Panorama: The rise of omnidirectional vision in the embodied ai era. *arXiv preprint*, 2026.

Yikang Zhou, Tao Zhang, Dizhe Zhang, Shunping Ji, Xiangtai Li, and Lu Qi. Dense360: Dense understanding from omnidirectional panoramas. *arXiv preprint arXiv:2506.14471*, 2025.

Appendix

Contents

A More Analytical Study

While the main text validates the efficacy of each proposed module via extensive ablation studies, this section provides further empirical analyses to comprehensively illustrate the findings from our experimental process.

A.1 Superparameter Analysis of Visual Prompt

Grid Prompting Style: Lines vs. Color Blocks We explore different visual prompting styles to effectively demarcate the spatial cells for the VLM. The primary comparison is between utilizing a grid composed of distinct solid lines versus applying semi-transparent colored blocks (color-coding each grid cell differently). A visualization is shown in Fig. 8. As detailed in Table 7 and Fig. 7, the line-based prompt achieves the optimal performance across all metrics, yielding a gIoU of 72.69 and a cIoU of 63.85. Conversely, applying color blocks significantly degrades performance. As the opacity level (α) of the color blocks increases from 50 to 150, the gIoU continuously drops from 70.91 to 68.67, and the cIoU drops from 59.75 to 57.46. This steady performance decline occurs because the color block overlay inevitably destroys the original color distribution of the underlying panoramic image. Such severe color distortion introduces unnatural tinting that confuses the VLM, severely interfering with its perception of the inherent texture, material, and color attributes critical for precise semantic deduction. In contrast, drawing simple grid lines minimally perturbs the original visual features of the scene, allowing the VLM to maintain robust recognition capability while successfully parsing the spatial layout.

Table 7 Performance comparison of different visual prompt types. Row 1 indicates the default setting used in our final pipeline.

#	Visual Prompt Type	gIoU↑	cIoU↑	P_{50} ↑	P_{50-95} ↑
1	Line	**72.69**	**63.85**	**76.29**	**66.13**
2	Color ($\alpha = 50$)	70.91	59.75	74.92	64.58
3	Color ($\alpha = 100$)	69.96	57.98	74.09	63.76
4	Color ($\alpha = 150$)	68.67	57.46	71.48	62.70

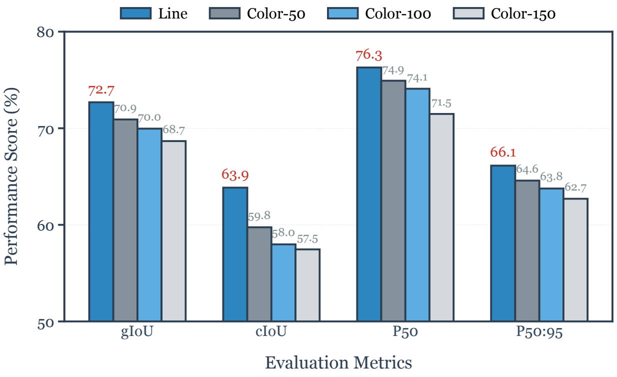
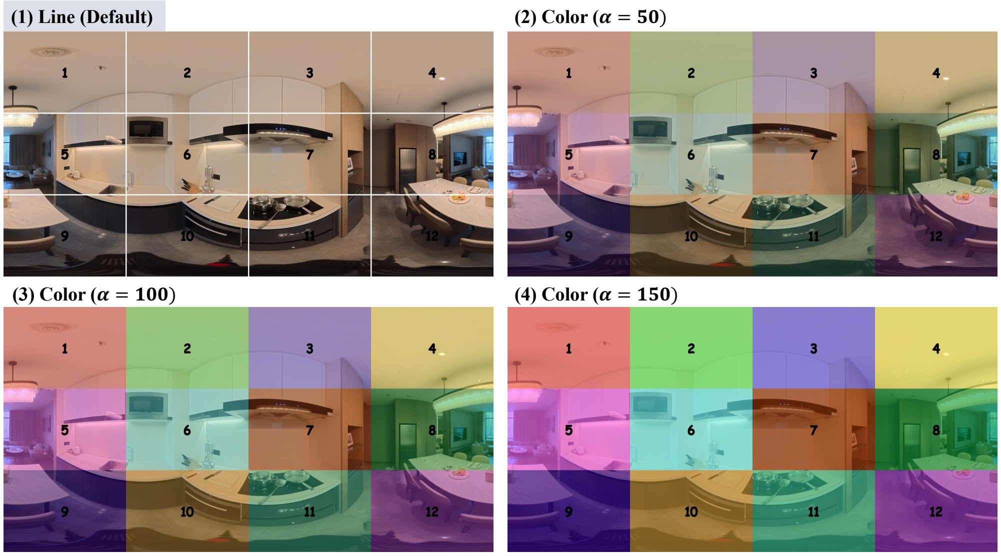


Figure 7 Performance comparison of different visual prompting types. Red highlights the best results for each metric.



(1) Line (Default)

(2) Color ($\alpha = 50$)

(3) Color ($\alpha = 100$)

(4) Color ($\alpha = 150$)

Figure 8 Visualization of Different Visual Prompt in Table 7.

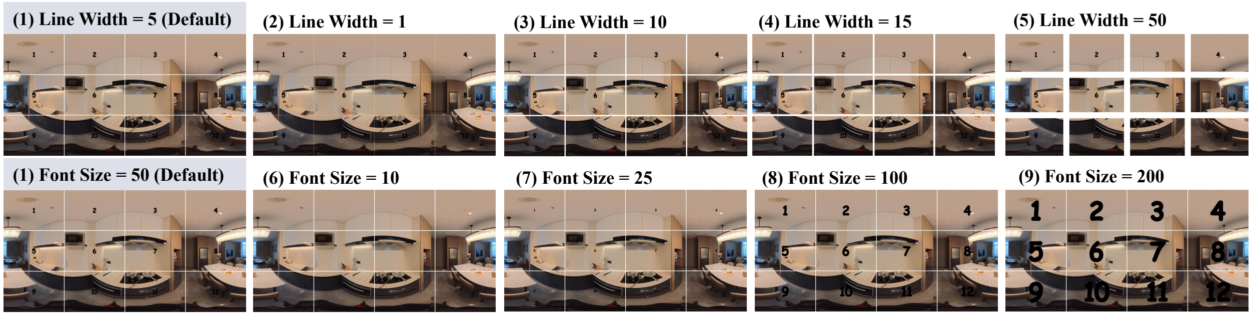


Figure 9 Visualization of different line width and font size in Table 8.

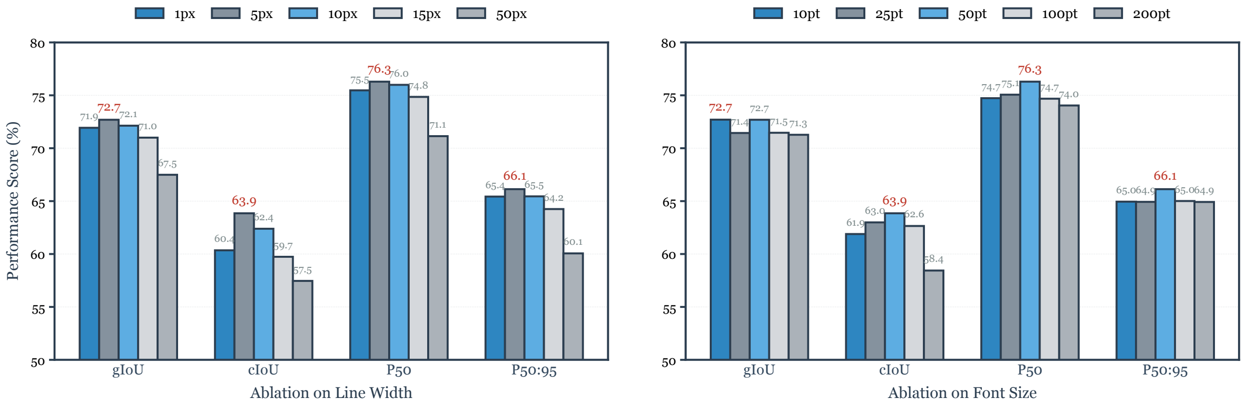


Figure 10 Ablation study on line width and font size. Red highlights the best results.

Grid Line Width We evaluate the impact of grid line width by drawing the visual prompt on downsampled panoramic images of 2000 × 1000 resolution. As shown in Table 8 and Fig. 10, our pipeline exhibits strong robustness to this parameter. Interestingly, even an extremely thin line of just 1 pixel (accounting for merely 0.1% of the image height) is highly effective in helping the VLM improve its spatial perception, yielding a competitive gIoU of 71.93 and cIoU of 60.35. The performance remains stable and optimal within a moderate range, peaking at a line width of 5 pixels. However, when the line width becomes excessively wide, such as 50 pixels (spanning 5% of the image height), the thick grid lines severely fragment the scene and occlude critical fine-grained visual information. This fragmentation causes a drastic performance drop, with gIoU falling to 67.49 and cIoU dropping to 57.46. Qualitative comparisons are provided in Fig. 9.

Table 8 Performance comparison of different line widths and font sizes. Row 1 indicates the default setting used in our final pipeline.

#	Line Width	Font Size	gIoU↑	cIoU↑	P_{50} ↑	P_{50-95} ↑
1	**5**	**50**	72.69	**63.85**	**76.29**	**66.13**
2	1	**50**	71.93	60.35	75.47	65.43
3	10	**50**	72.13	62.39	75.99	65.45
4	15	**50**	71.00	59.74	74.85	64.25
5	50	**50**	67.49	57.46	71.14	60.06
6	**5**	10	**72.70**	61.89	74.73	64.95
7	**5**	25	71.44	62.99	75.06	64.93
8	**5**	100	71.46	62.65	74.68	65.01
9	**5**	200	71.27	58.44	74.04	64.92

Font Size of Grid Coordinates Similarly, our pipeline demonstrates strong robustness to the font size of the numerical indices (1 to 12). As reported in Table 8, varying the font size within a broad range (from 10 to 100) yields consistently stable results, with an intermediate font size of 50 (occupying about 5% of the image height) achieving the optimal balance. However, akin to the line width degradation, excessively large text negatively impacts performance. For instance, when the font size is increased to 200 (spanning 20% of the image height), the oversized numbers act as severe artificial occlusions. These massive digits mask important contextual cues and target objects in the underlying image, causing the cIoU to drop notably to

Table 9 Performance comparison of different grid resolutions. Row 3 indicates the default setting used in our final pipeline.

#	Grid Resolution	gIoU↑	cIoU↑	P_{50} ↑	P_{50-95} ↑
1	2 × 2	61.45	52.66	64.14	53.50
2	3 × 3	72.20	62.68	75.91	66.08
3	**4 × 3**	**72.69**	**63.85**	**76.29**	**66.13**
4	4 × 4	71.91	60.15	76.27	65.59
5	5 × 5	72.11	63.10	75.93	65.25
6	10 × 10	60.07	61.66	62.11	53.16

Table 10 Performance comparison of PAP using different VLM models. Row 2 indicates the default setting used in our final pipeline.

#	VLM Model	gIoU↑	cIoU↑	P_{50} ↑	P_{50-95} ↑
1	Gemini-3-Flash	74.03	67.65	77.74	67.20
2	Qwen-3-VL-32B	72.69	63.85	76.29	66.13
3	Qwen-3-VL-8B	68.93	55.57	72.43	62.34
4	Qwen-3-VL-4B	66.28	52.79	70.15	59.94
5	Qwen-2.5-VL-7B	65.47	52.12	68.08	57.35

58.44. The visual impact of varying font sizes is also illustrated in Fig. 9.

Analysis of Grid Resolution. Furthermore, we analyze the impact of different visual grid resolutions. As shown in Table 9 and Fig. 11, the performance remains robust across a moderate range of resolutions, with 3 × 3, 4 × 3, 4 × 4, and 5 × 5 grids yielding comparable results. However, both overly coarse and excessively dense grids negatively impact performance. For instance, a 2 × 2 grid provides insufficient spatial constraints, offering limited assistance to the VLM. Conversely, a dense 10 × 10 grid containing 100 cells imposes a heavy burden on the VLM's reasoning capacity, leading to a noticeable performance drop. We adopt the 4 × 3 grid as our default setting since it achieves the best performance.

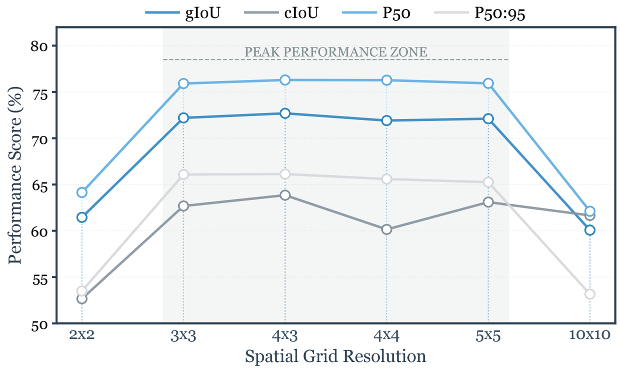


Figure 11 Comparison between different grid resolutions.

A.2 Analysis of Different VLM backbones

To evaluate the impact of the Vision-Language Model (VLM) backbone on our pipeline, we conduct experiments across models with varying capabilities and scales, as presented in Table 10. First, we observe a strong positive correlation between the overall performance metrics and the intrinsic capabilities of the VLM. Specifically, substituting our default setting, Qwen-3-VL-32B, with a more powerful, closed-source model (Gemini-3-Flash) yields substantial improvements across all metrics. Conversely, adopting smaller open-source models (such as Qwen-3-VL-8B/4B or Qwen-2.5-VL-7B) results in a corresponding decline in performance.

Furthermore, these results demonstrate the ***high flexibility*** of our proposed pipeline. It can seamlessly integrate state-of-the-art closed-source models to pursue absolute performance, effectively serving as a robust offline data generation engine. Alternatively, for scenarios where computational efficiency is a priority, the pipeline can readily adopt more lightweight models to maintain a desirable balance between performance and inference cost.

A.3 Analysis of Different Resolution

Table 11 Performance comparison of different image resolutions. Row 2 indicates the default resolution used in our final pipeline.

#	Resolution	gIoU↑	cIoU↑	P_{50} ↑	P_{50-95} ↑
1	4000 × 2000	**73.32**	58.98	**76.81**	**67.30**
2	2000 × 1000	72.69	**63.85**	76.29	66.13
3	1000 × 500	66.46	56.45	70.78	59.18

To evaluate the impact of image resolution, we conducted experiments across three different settings: 4000 × 2000, 2000 × 1000, and 1000 × 500. As shown in Table 11, the model exhibits relatively poor performance at 1000 × 500. This is likely due to the low resolution preventing the model from capturing sufficient fine-grained details. Conversely, increasing the resolution to 2000 × 1000 yields a significant performance boost. This aligns with our expectations, as 2000 × 1000 is visibly much sharper to the naked eye, providing adequate clarity for detail extraction. However, further scaling the resolution to 4000 × 2000 results in marginal gains; while metrics such as gIoU, P_{50}, and P_{50-95} show slight improvements, cIoU actually declines. We attribute this to our Recursive Visual Routing mechanism, which effectively leverages a coarse-to-fine approach. By initially analyzing global context from a highly downsampled image and subsequently narrowing the field of view with reduced downsampling to capture finer details, the 2000 × 1000 resolution proves to be entirely sufficient for this paradigm. Furthermore, since the 4000 × 2000 resolution introduces four times the number of image tokens compared to 2000 × 1000, the inference latency is effectively doubled. Therefore, to achieve an optimal trade-off between performance and inference efficiency, we adopted 2000 × 1000 as our default configuration.

B Discussion

B.1 Acceptable but Unimpressive Latency

As demonstrated in Table 2 in the main text, our PAP significantly outperforms other baselines while maintaining comparable inference latency(*∼10 seconds*). However, for real-world robotic deployment, minimizing latency is always desirable. As shown in Table 10, we explored using more lightweight VLMs, such as Qwen3-VL-4B, as the backbone. Although this resulted in an approximate *8%* performance drop, it still vastly exceeds the baseline methods and bounds the inference latency to *2–3 seconds*. We consider this latency acceptable for the high-level planning task of locating target objects within a massive and complex 360-degree environment based on intricate instructions. For future work, we plan to further explore latency reduction; for instance, we could leverage our proposed PAP for offline data generation to distill an smaller end-to-end model, thereby further minimizing inference time.

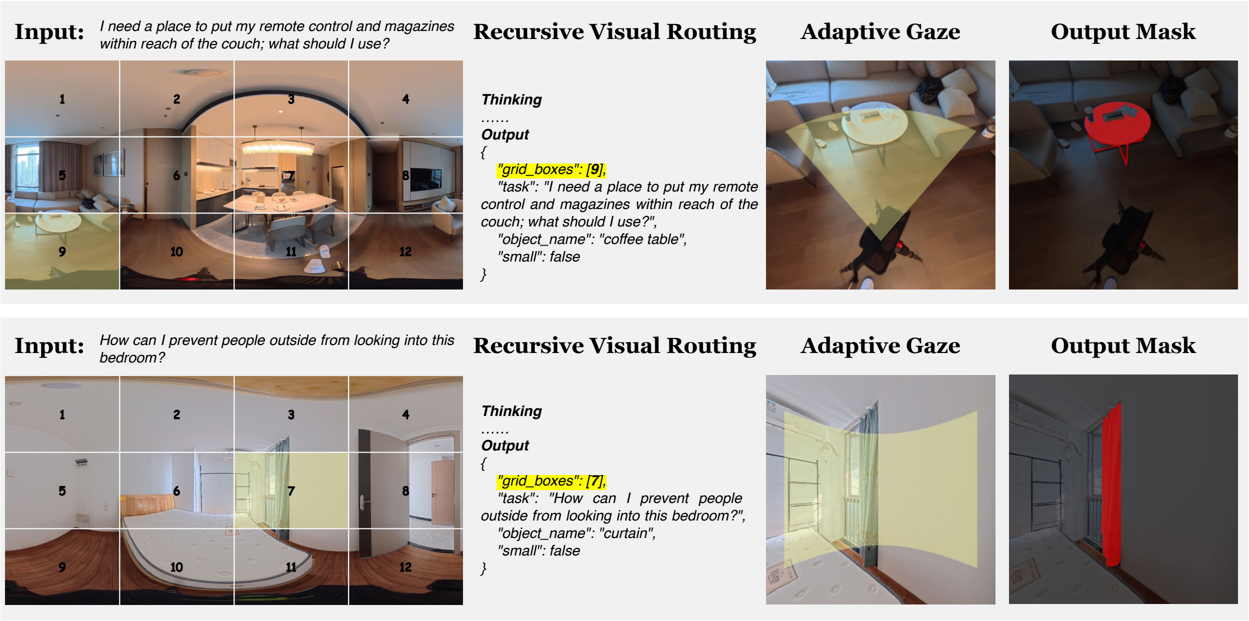
B.2 End-to-End Model v.s. Decoupled Pipeline

Following the success of prior work (Zhang et al., 2025c) that decouples the reasoning and grounding stages of affordance prediction, the proposed PAP adopts a similar paradigm. Compared to end-to-end solutions, the decoupled approach offers several key advantages: 1) greater flexibility in integrating various Vision-Language Models (VLMs); 2) the ability to fully exploit the specific strengths of individual models; 3) strong zero-shot generalization capabilities without the need for training on massive datasets; and 4) enhanced interpretability, as the intermediate outputs of each step are accessible, which facilitates debugging and optimization. As the first exploration of the panoramic affordance prediction task, our work embraces this decoupled strategy. The state-of-the-art performance of PAP on PAP-12K demonstrates the effectiveness of this decoupled philosophy within the panoramic domain.

Nevertheless, we recognize the significant potential of end-to-end models for this task. While we have introduced several tailored designs to bridge the domain gap between panoramic and conventional camera imaging, a model capable of natively understanding Equirectangular Projection (ERP) images and efficiently inferring affordances would offer even greater value for downstream applications. Notably, our proposed framework intrinsically serves as a robust, zero-shot data generation engine. In future work, PAP can be utilized to further scale up the training data, paving the way for the highly efficient end-to-end models.

B.3 Managing Cascading Errors

While decoupled models offer numerous advantages over end-to-end models, a potential pitfall is the accumulation of errors. We accounted for this in our design, primarily by managing the errors generated by our Recursive Visual Routing module. We observed that the model outputs accurate grid indices in the vast majority of cases, but it is prone to minor errors in one specific scenario: when a small portion of an object happens to cross a grid boundary. In this situation, the model yields one of two possible outputs: *1)* It accurately provides the indices of both grids the object spans. This occurs most of the time and poses no issue, allowing for direct progression to the next step. *2)* It outputs only one of the two grids. When this happens and the process advances to the adaptive gaze step, if the selected FoV (Field of View) only covers this single grid exactly, the object is very likely to be split in half, which compromises the subsequent grounding.


Input:
I need a place to put my remote control and magazines within reach of the couch; what should I use?
Recursive Visual Routing
Adaptive Gaze
Output Mask
1
2
3
4
5
6
7
8
9
10
11
12
Input:
How can I prevent people outside from looking into this bedroom?
Recursive Visual Routing
Adaptive Gaze
Output Mask
1
2
3
4
5
6
7
8
9
10
11
12

```
Thinking
......
Output
{
    "grid_boxes": [9],
    "task": "I need a place to put my remote control and magazines within reach of the couch; what should I use?",
    "object_name": "coffee table",
    "small": false
}
```

```
Thinking
......
Output
{
    "grid_boxes": [7],
    "task": "How can I prevent people outside from looking into this bedroom?",
    "object_name": "curtain",
    "small": false
}
```

Figure 12 Visualization of the error cases of Recursive Visual Routing. We also visualize the coverage of the grid, where the yellow translucent areas show the areas covered by the grid in the ERP view and the FoV view. It can be seen that our Adaptive Gaze module effectively covers the entire grid containing the object, with a small margin of redundancy.

To effectively resolve this error, we proposed a solution: in the adaptive gaze module, we add a margin of redundancy (around 10 degrees) to the adaptive gaze's FoV. This way, when reverting to the normal view, the FoV covers a small area beyond the grid, ensuring that objects situated right at the boundary between two grids are not split. As shown in two cases in Fig. 12, the correct objects (coffee table in the top and curtain in the bottom) have an extremely small portion falling outside the grid. During the RVR process, the VLM fails to correctly output all the grids occupied by the object. However, our Adaptive Gaze module, by adding a small margin to the FoV, successfully resolves this error and allows the entire pipeline to produce the correct mask.

C More Implementation Details

C.1 System Prompt of Recursive Visual Routing

System Prompt: Recursive Visual Routing

Environment Setting

Given a panoramic image of a scene, the task is to decide which object to use and predict the part of the object that matches the provided task. The task instruction is "TASK". To assist your spatial reasoning, the image is overlaid with a *4x3 grid marked with large numbers 1 through 12.* The grid follows a standard reading order: Top-Row (1, 2, 3, 4), Middle-Row (5, 6, 7, 8), and Bottom-Row (9, 10, 11, 12). These numbers serve as your spatial reference system.

Your task

1. Identify the target object in the 4x3 grid panoramic image according to the task instructions.
2. Accurately locate the grid boxes corresponding to the target object.
3. *STRICTLY* follow the output format, especially the JSON format.

Follow these reasoning steps:

1. *Step 1 Identify the Target*
   - First identify the target to use from the scene that best matches the provided task.
   - Then, identify the key components of this object in the image (e.g., shape, features, possible points of interaction).
   - Finally, analyze the object and the task instruction, provide the final answer "object_name" or a more specific part of the object "object_part".
   - *Rule*: If there are multiple similar objects in the image, please make sure your "object_name" or "object_part" is uniquely identifiable and clearly distinguishable from the other similar objects by adding additional descriptions.

2. *Step 2 Spatial Occupancy (Grid Mapping)*
   - Closely observe the *4x3 grid numbering (1-12)*.
   - List *ALL* grid boxes that contain any part of the target object.
   - *Rule*: If the object's body or edges cross the edge line, include all grid boxes on both sides of the line.
3. *Step 3 Clarity Analysis (Small Object Refinement)*
   - When the target object occupies *only ONE* grid box, judge whether the object appears *small*.
   - *Rule*: An object is "small" compared to other normal objects, making it difficult to precisely segment for the task.
4. *Step 4 Output*
   - Output the result in a structured JSON format. *STRICTLY* follow the output format.

Output format:

Thinking Thinking process 1. Identify the target object. 2. Identify the number of grid boxes. 3. Identify whether the target object is small.

Output { "grid_boxes": [index1, index2, . . .], // e.g., [4, 5] or [5] — first-level 4×3 grid indices (1–12) "task": "the task instruction", "object_name": "the name of the object", "object_part": "the part of the object", // if the whole object is the target, leave it the same as "object_name" "small": true/false }

C.2 Formulation of the Adaptive Gaze Module

To extract a distortion-free rectilinear image (viewport) from an Equirectangular Projection (ERP) panorama, an inverse mapping approach is employed in our Adaptive Gaze Module. This approach guarantees a dense target image without holes. The procedure involves mapping each pixel of the target viewport to a 3D ray, applying rotational transformations based on the desired viewing direction, and finally mapping the ray back to the ERP image coordinates to sample the pixel values.

Let the target rectilinear image have a width of W, a height of H, and a horizontal Field of View denoted as FOV. The focal length f of the virtual camera is first calculated as:

$$f = \frac{W}{2\tan(\mathrm{FOV}/2)} \tag{1}$$

For each pixel (x, y) in the target image, assuming the optical center is at (c_x, c_y) (typically the center of the image), its local 3D ray coordinates (X_c, Y_c, Z_c) are defined as:

$$X_c = x - c_x \tag{2}$$

$$Y_c = y - c_y \tag{3}$$

$$Z_c = f \tag{4}$$

To simulate the camera orientation, these 3D rays are rotated according to the specified yaw (θ) and pitch (φ) angles. Let the rotation matrices for pitch (around the X-axis) and yaw (around the Y-axis) be R_x and R_y, respectively. The world coordinates (X_w, Y_w, Z_w) of the ray are obtained by applying these rotations:

$$\begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} = R_y R_x \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \tag{5}$$

The rotated 3D direction vectors in the world coordinate system are then converted into spherical coordinates, representing longitude λ and latitude ϕ:

$$\lambda = \arctan 2(X_w, Z_w) \tag{6}$$

$$\phi = \arcsin\left(\frac{Y_w}{\sqrt{X_w^2 + Y_w^2 + Z_w^2}}\right) \tag{7}$$

where $\lambda \in [-\pi, \pi]$ and $\phi \in [-\pi/2, \pi/2]$.

Finally, the spherical coordinates are normalized and mapped to the 2D pixel coordinates (U, V) of the original ERP image, which has a width of W_{erp} and a height of H_{erp}:

$$U = \left(\frac{\lambda}{2\pi} + 0.5\right) \times W_{\text{erp}} \tag{8}$$

$$V = \left(\frac{\phi}{\pi} + 0.5\right) \times H_{\text{erp}} \tag{9}$$

Once the continuous coordinates (U, V) are computed for all pixels in the target viewport, bilinear interpolation is applied to sample the corresponding color values from the discrete grid of the ERP image, thereby constructing the final rectilinear output.

C.3 Details of the Implementation of Baseline Methods

As we are the first to tackle the panoramic affordance prediction task, there are no existing baseline methods can directly process the PAP-12K. All of the baseline methods are designed for single perspective image. Therefore, we kindly introduce specific adaptations for them to ensure a fair and valid comparison. First, methods such as Affordance-R1, AffordanceVLM, VisionReasoner, LISA, and OV-Seg are trained exclusively on standard perspective image affordance datasets. In our preliminary experiments, directly feeding them full-resolution panoramic images resulted in near-complete failure, yielding performance metrics close to zero. To ensure meaningful evaluations, we resize the panoramic input images to match the respective training resolutions of each method before testing. On the other hand, A4-Agent is a training-free framework whose underlying Vision-Language Model (VLM) inherently possesses some capability to process ultra-high-resolution images. Thus, while it does not fail completely when given full-resolution inputs, its performance remains suboptimal and incurs prohibitive computational costs. Consequently, we resize the input images for A4-Agent to 2000 × 1000. This resolution precisely matches the input resolution used in the initial stage of our proposed pipeline, further ensuring the validity and fairness of the evaluation. All the experiments and the calculation of metrics in this paper are conducted on the same Server with A40 GPU. For the baseline methods, we use the official code and hyperparameters provided by the authors, except for the resize operation mentioned above.

D More Details about the Data Annotation Process

Here, we provide further details regarding the data annotation process for Affordance VQA. Furthermore, to maximize our contribution to the research community, we will ***fully open-source*** not only the PAP-12K and PAP themselves, but also our complete data annotation pipeline. This encompasses the Agent employed in Phase 1 and the customized WebUI annotation tool developed for Phase 2.

Phase 1: Affordance Question Formulation. First, regarding question generation, to facilitate scalable production, we adopted a pipeline consisting of batch generation by an AI agent, followed by manual human refinement and filtering. For this agent, we utilized the system prompt illustrated in Fig. D. Our design process was guided by the following core principles:

1. Due to the distortion present at the poles of ERP images, we implemented a specific processing step: we simultaneously feed the original image along with the converted cubemap projections—totaling seven images—into the VLM to generate the questions.

2. Instead of visual identification ("Find the cup"), the AI must deduce the functional purpose of an object and frame the question as a human need (e.g., "I need something to hold water"). Moreover, the target object must serve as the absolute unique solution to that specific need within the room.

3. The prompt mandates strict boundaries to prevent "technically correct but ambiguous" VQA pairs. It forces the Agent to actively distinguish between subtle functional roles, such as the difference between a surface and a tool (e.g., a whiteboard vs. a marker) or a container and its contents (e.g., a cup vs. water).

4. To ensure question diversity, we instructed the Agent to formulate 2 to 4 distinct questions for the same target object based on its various functional uses.

To further ensure the quality and diversity of the generated questions, we employed two of the most powerful closed-source models (Gemini-3-Pro and GPT-5) as our backbones to independently conduct batch generation. Ultimately, their respective outputs were merged, followed by screening and refinement by human experts.

System Prompt: Affordance Question Generation

Role
You are an expert AI agent specializing in Visual Question Answering (VQA) dataset generation.

Task
You will be provided with **7 images** for each scene: 1. **Image 1:** A 360-degree equirectangular panoramic view of the room. 2. **Images 2-7:** Six cubemap projection faces generated from the panoramic view (in order: Front, Right, Back, Left, Top, Bottom), these images can assist you to understand the spatial context and object details.
Your task is to analyze these images collectively to identify distinct objects and generate "Affordance VQA" pairs. You must prioritize generating unambiguous questions where the target object is the only logical answer.

Definition of Affordance
An "affordance" defines the possible actions an agent can perform with an object. * **Goal:** Create a query regarding a need/intent where the **unique solution** is the specific object.

Step-by-Step Instructions

1. **Object Detection & Adaptive Description:** Identify objects and valid interactable parts using all available views.
   - **Rule: Adaptive Detail.**
     - **Simple Scenes/Unique Objects:** If an object is unique and unambiguous in the scene (e.g., only one sofa), use a **concise name** (e.g., "sofa").
     - **Complex Scenes/Ambiguous Objects:** If there are multiple similar objects or the scene is crowded, you **MUST** provide a **detailed description** including **Location** (e.g., "on the left desk"), **Appearance** (e.g., "red ceramic"), or **State** (e.g., "open/closed") to distinguish it.
   - *Goal:* Keep it short when possible, but specific when necessary.
2. **Disambiguation Logic:** Ensure the question points *only* to the target object.
   - *Distinguish Surface vs. Tool:* "Write **on**" (Whiteboard) vs. "Write **with**" (Pen).
   - *Distinguish Container vs. Content:* "Pour **into**" (Cup) vs. "Drink" (Water).
   - *Distinguish Identical Objects:* Use the detailed description from Step 1 if multiple instances exist (e.g., "the open laptop" vs "the closed laptop").
3. **Question Phrasing (CRITICAL):**
   - **DO NOT** start every question with "I need" or "I want".
   - You **MUST** rotate between the following **4 phrasing styles**:
     - **Style A (Search/Location):** "Where can I find a [property] place to [action]?"
     - **Style B (Problem/Context):** "[Situation description], what can I use to [solve it]?"
     - **Style C (Operational/How-to):** "How can I [action] using an object in this room?"
     - **Style D (Functional Selection):** "Which object allows me to [specific capability]?"
4. **Formatting:** Output strictly in **JSON Lines** format.

Output Format Schema
Each line must follow this exact JSON structure: {"object": "<Adaptive Object Description (Concise or Detailed)>", "affordance_question": ""}

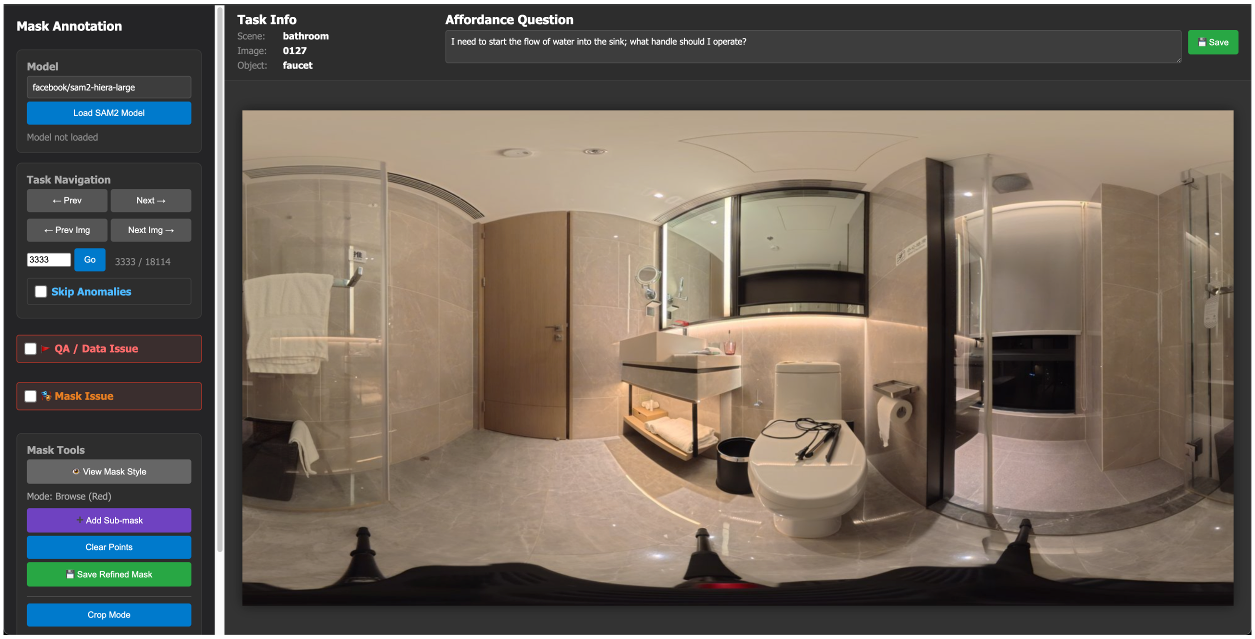
Examples (Demonstrating Variety & Adaptive Detail)
Input Image Context: A living room with a beige sofa, a wall-mounted TV, drawn curtains, a wooden coffee table, and TWO remote controls (one on the sofa, one on the table).
Output: {"object": "sofa", "affordance_question": "I am feeling exhausted and looking for a soft place to lie down."} {"object": "television", "affordance_question": "Which device should I look at to watch the evening news?"} {"object": "curtains", "affordance_question": "The sun is too bright in my eyes; what can I use to block the light?"} {"object": "coffee_table", "affordance_question": "Where is a stable surface specifically designed to hold my beverages?"} {"object": "remote control on the sofa armrest", "affordance_question": "How can I change the channel without standing up?"} {"object": "carpet", "affordance_question": "I want to

sit on the floor but need something warmer than the tiles."}

Constraints

1. **Variety is Key:** Ensure the questions sound natural and distinct from one another. Avoid repetitive sentence starters.
2. **Avoid Ambiguity:** The question should not arguably apply to other common objects.
3. **Adaptive Detail:** Use detailed descriptions ONLY when necessary to distinguish objects. Otherwise, keep the object name concise.
4. **Format:** One JSON object per line. No markdown blocks.
5. **Language:** English only.

Phase 2: Mask Segmentation. Once the affordance question-answer pairs were established, the subsequent step involved generating the corresponding segmentation masks based on the answers. This process was manually executed by a dedicated annotation team of five members. To streamline the workflow, we developed a customized WebUI specifically for this dataset, as illustrated in Fig. 13. The front-end of the interface displays relevant information to the annotator, such as the affordance question and answer, while the back-end integrates the SAM2 model, enabling users to provide prompts for accurate, interactive segmentation. The entire annotation effort spanned two months, ultimately yielding over 15,000 affordance VQA instances paired with precise segmentation masks.



Mask Annotation

Model: facebook/sam2-hiera-large

Model not loaded

3333 / 18114

Mode: Browse (Red)

Task Info

Scene: bathroom

Image: 0127

Object: faucet

Affordance Question

I need to start the flow of water into the sink; what handle should I operate?

Figure 13 The WebUI for Affordance VQA Annotation.

Phase 3: Final Verification. The final step involves a comprehensive review of the generated question-answer pairs and segmentation masks to ensure high quality and eliminate ambiguity. Specifically, the annotation team conducted a rigorous cross-checking process across the entire dataset. Following two complete rounds of manual review, we filtered out over 1,000 potentially ambiguous instances from the initial pool of 15,000+ samples. This rigorous verification process resulted in a final, high-quality dataset consisting of 13,943 affordance VQA instances paired with segmentation masks.

E More Visualizations of PAP-12K

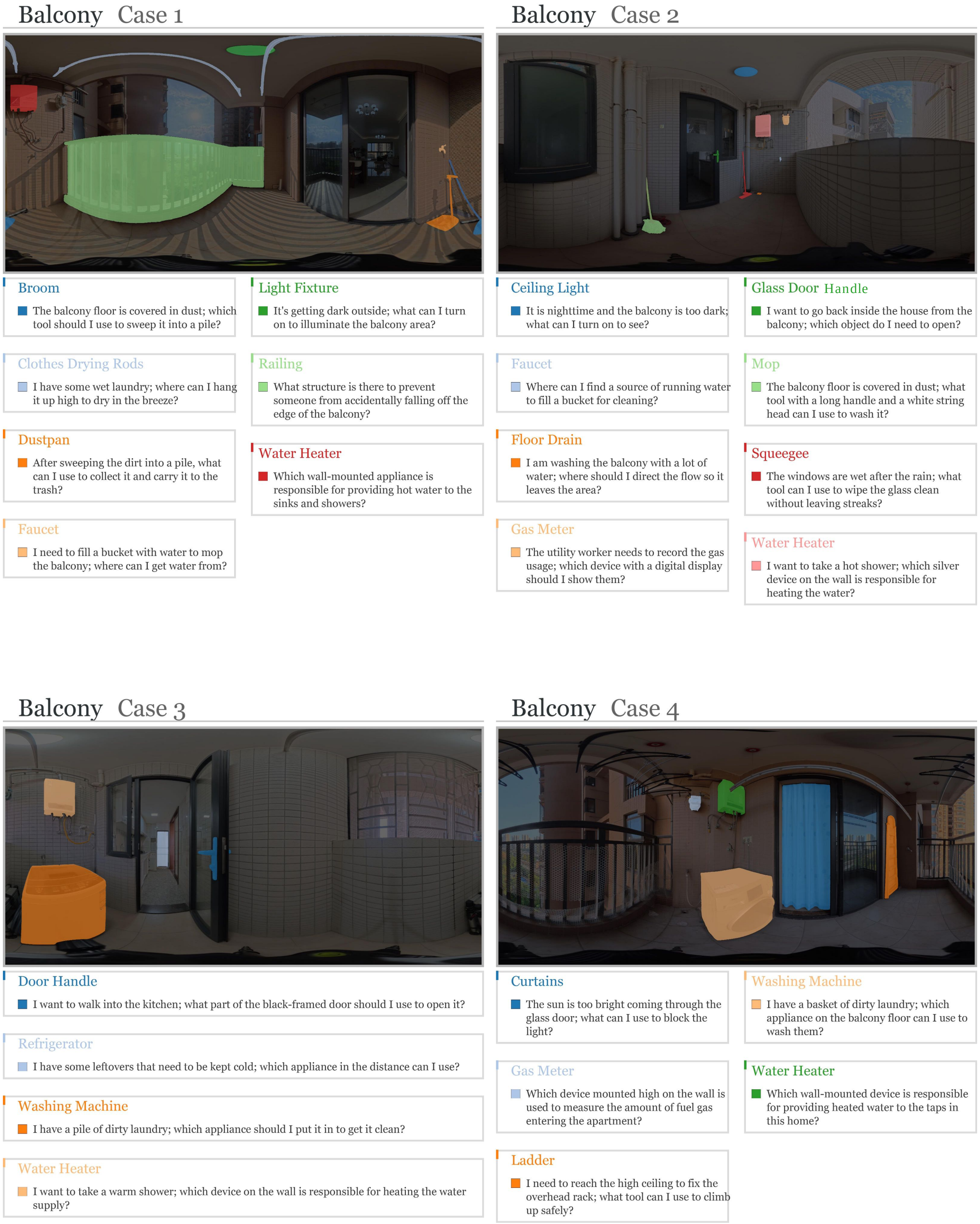
We provide more visualizations of the PAP-12K in this section. For each scene type, we randomly select 4 cases. Each case features a variable number of objects with their corresponding ground-truth affordance maps and questions. All of them are visualized in distinct colors within the same panoramic image. Please refer to Fig. 14 (Balcony), Fig. 15 (Bathroom), Fig. 16 (Bedroom), Fig. 17 (Classroom), Fig. 18 (Corridor), Fig. 19 (Gym), Fig. 20 (Kitchen), Fig. 21 (Livingroom), Fig. 22 (Office), Fig. 23 (Pantry), Fig. 24 (Workshop), and Fig. 25 (Others) for detailed visualizations.

F More Qualitative Comparisons on PAP-12K

Here, we present further qualitative comparisons between our proposed PAP and the baseline methods evaluated on PAP-12K. Detailed visualizations can be found in Figs. 26–37. As illustrated, our approach consistently demonstrates superior performance over the baseline methods across all scene categories.

G Ethics Statement

The data collection process for this research involved capturing panoramic images in both private and public environments. To ensure strict adherence to ethical standards and privacy protection, the following protocols were implemented: **1) Data Collection Consent:** For specific private or semi-public locations, explicit permission was obtained from the respective property owners or managers prior to capturing the images. Images in public areas were captured without disrupting normal public activities. **2) Privacy Protection:** During the image acquisition phase, we proactively avoided capturing highly private or sensitive areas. Furthermore, we deliberately ensured that no frontal human faces or any identifiable personal features were captured during the process.



Balcony Case 1

Broom
🟦 The balcony floor is covered in dust; which tool should I use to sweep it into a pile?

Clothes Drying Rods
🟦 I have some wet laundry; where can I hang it up high to dry in the breeze?

Dustpan
🟧 After sweeping the dirt into a pile, what can I use to collect it and carry it to the trash?

Faucet
🟧 I need to fill a bucket with water to mop the balcony; where can I get water from?

Light Fixture
🟩 It's getting dark outside; what can I turn on to illuminate the balcony area?

Railing
🟩 What structure is there to prevent someone from accidentally falling off the edge of the balcony?

Water Heater
🟥 Which wall-mounted appliance is responsible for providing hot water to the sinks and showers?

Balcony Case 2

Ceiling Light
🟦 It is nighttime and the balcony is too dark; what can I turn on to see?

Faucet
🟦 Where can I find a source of running water to fill a bucket for cleaning?

Floor Drain
🟧 I am washing the balcony with a lot of water; where should I direct the flow so it leaves the area?

Gas Meter
🟧 The utility worker needs to record the gas usage; which device with a digital display should I show them?

Glass Door Handle
🟩 I want to go back inside the house from the balcony; which object do I need to open?

Mop
🟩 The balcony floor is covered in dust; what tool with a long handle and a white string head can I use to wash it?

Squeegee
🟥 The windows are wet after the rain; what tool can I use to wipe the glass clean without leaving streaks?

Water Heater
🟥 I want to take a hot shower; which silver device on the wall is responsible for heating the water?

Balcony Case 3

Door Handle
🟦 I want to walk into the kitchen; what part of the black-framed door should I use to open it?

Refrigerator
🟦 I have some leftovers that need to be kept cold; which appliance in the distance can I use?

Washing Machine
🟧 I have a pile of dirty laundry; which appliance should I put it in to get it clean?

Water Heater
🟧 I want to take a warm shower; which device on the wall is responsible for heating the water supply?

Balcony Case 4

Curtains
🟦 The sun is too bright coming through the glass door; what can I use to block the light?

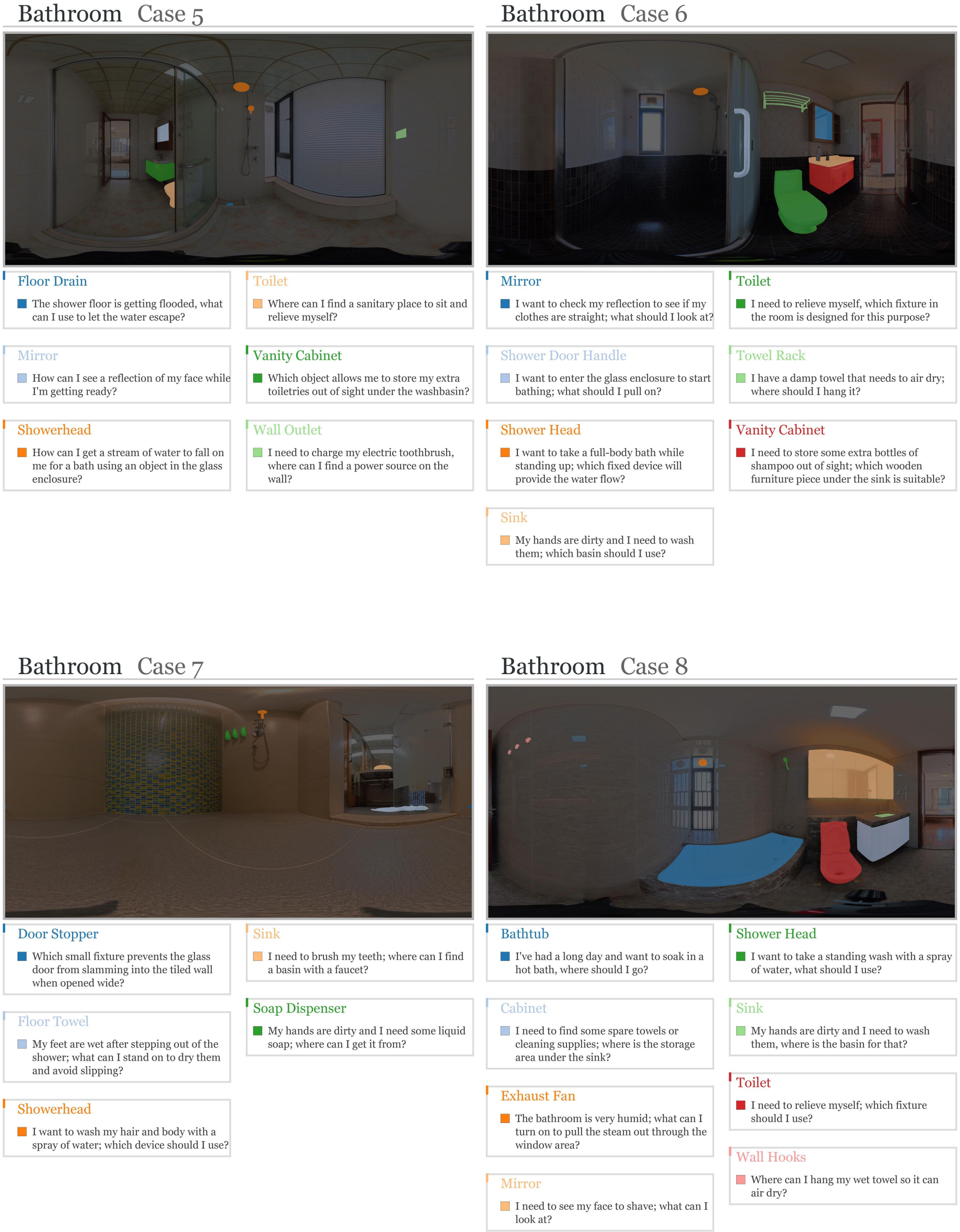
Gas Meter
🟦 Which device mounted high on the wall is used to measure the amount of fuel gas entering the apartment?

Ladder
🟧 I need to reach the high ceiling to fix the overhead rack; what tool can I use to climb up safely?

Washing Machine
🟧 I have a basket of dirty laundry; which appliance on the balcony floor can I use to wash them?

Water Heater
🟩 Which wall-mounted device is responsible for providing heated water to the taps in this home?

Figure 14 Visualizations of PAP-12K(Balcony).



Bathroom Case 5

Floor Drain

🟦 The shower floor is getting flooded, what can I use to let the water escape?

Toilet

🟧 Where can I find a sanitary place to sit and relieve myself?

Mirror

🟦 How can I see a reflection of my face while I'm getting ready?

Vanity Cabinet

🟩 Which object allows me to store my extra toiletries out of sight under the washbasin?

Showerhead

🟧 How can I get a stream of water to fall on me for a bath using an object in the glass enclosure?

Wall Outlet

🟩 I need to charge my electric toothbrush, where can I find a power source on the wall?

Bathroom Case 6

Mirror

🟦 I want to check my reflection to see if my clothes are straight; what should I look at?

Toilet

🟩 I need to relieve myself, which fixture in the room is designed for this purpose?

Shower Door Handle

🟦 I want to enter the glass enclosure to start bathing; what should I pull on?

Towel Rack

🟩 I have a damp towel that needs to air dry; where should I hang it?

Shower Head

🟧 I want to take a full-body bath while standing up; which fixed device will provide the water flow?

Vanity Cabinet

🟥 I need to store some extra bottles of shampoo out of sight; which wooden furniture piece under the sink is suitable?

Sink

🟧 My hands are dirty and I need to wash them; which basin should I use?

Bathroom Case 7

Door Stopper

🟦 Which small fixture prevents the glass door from slamming into the tiled wall when opened wide?

Sink

🟧 I need to brush my teeth; where can I find a basin with a faucet?

Floor Towel

🟦 My feet are wet after stepping out of the shower; what can I stand on to dry them and avoid slipping?

Soap Dispenser

🟩 My hands are dirty and I need some liquid soap; where can I get it from?

Showerhead

🟧 I want to wash my hair and body with a spray of water; which device should I use?

Bathroom Case 8

Bathtub

🟦 I've had a long day and want to soak in a hot bath, where should I go?

Shower Head

🟩 I want to take a standing wash with a spray of water, what should I use?

Cabinet

🟦 I need to find some spare towels or cleaning supplies; where is the storage area under the sink?

Sink

🟩 My hands are dirty and I need to wash them, where is the basin for that?

Exhaust Fan

🟧 The bathroom is very humid; what can I turn on to pull the steam out through the window area?

Toilet

🟥 I need to relieve myself; which fixture should I use?

Mirror

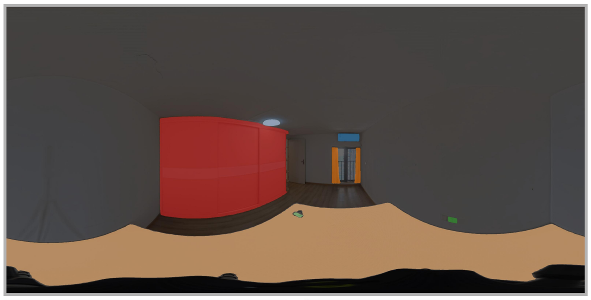
🟧 I need to see my face to shave; what can I look at?

Wall Hooks

🟥 Where can I hang my wet towel so it can air dry?

Figure 15 Visualizations of PAP-12K(Bathroom).

Bedroom Case 9



Air Conditioner

🟦 The room is getting too hot; which wall-mounted device can I use to cool the air?

Ceiling Light

🟦 It's getting dark outside; what can I turn on to illuminate the entire room?

Curtains

🟧 The morning sun is too bright; what can I pull across the window to darken the room?

Mattress

🟧 I'm feeling very sleepy and need a large, soft surface to lie down on.

Power Outlet

🟩 My phone battery is low; where can I plug in my charger on the wall?

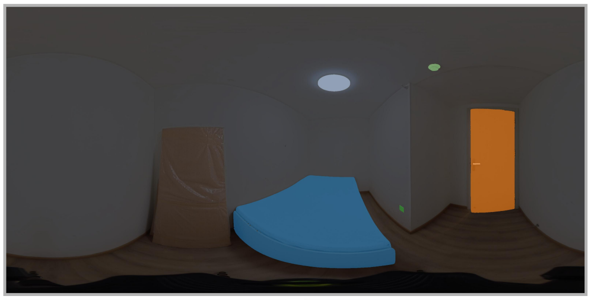
Remote Control

🟩 I want to adjust the temperature of the air conditioner from the bed; what should I pick up from the mattress?

Wardrobe

🟥 I have many shirts that need to be hung up; where can I store them?

Bedroom Case 10



Bed

🟦 I've had a long day and need to sleep; which object in the room is designed for me to lie down on?

Ceiling Light

🟦 Which fixture on the ceiling is currently providing light to the room?

Door

🟧 I want to leave this room; which large wooden object should I move to walk out?

Door Handle

🟧 What part of the door should I grasp and turn to open it?

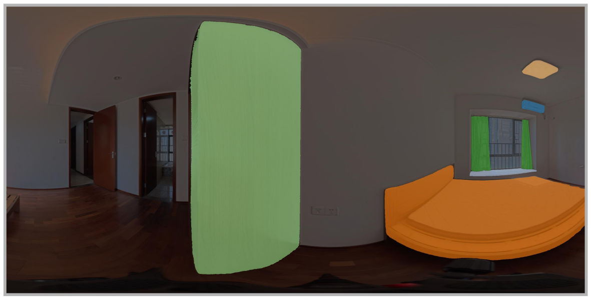
Power Outlet

🟩 My phone battery is low; where can I plug in my charger on the wall?

Smoke Detector

🟩 Which small device on the ceiling is responsible for alerting me in case of a fire?

Bedroom Case 11



Air Conditioner

🟦 The room is getting too hot; what device can I use to lower the temperature?

Bay Window Ledge

🟦 Where is a flat surface near the window where I could place a small potted plant or some books?

Bed

🟧 I am very tired and need a place to sleep comfortably for the night.

Ceiling Light

🟧 It's getting dark outside; what can I turn on to brighten up the entire room?

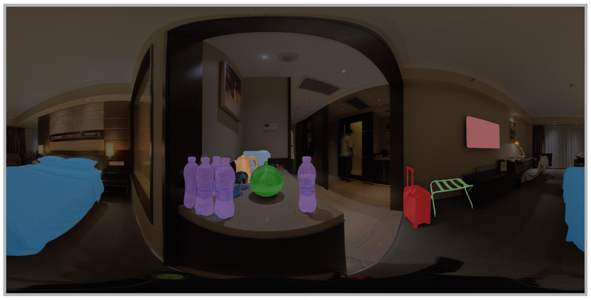
Curtains

🟩 The morning sun is too bright; what can I pull across to shade the room?

Wardrobe

🟩 I have many clothes that need to be hung up or stored away; where should I put them?

Bedroom Case 12



Bed

🟦 I am very tired and need to sleep; where is the best place to do that?

Coffee Maker

🟦 I need a caffeine boost in the morning; which machine on the counter can brew a cup of coffee for me?

Desk Lamp

🟧 It's getting dark and I need to read some documents at the desk; what can I turn on for focused light?

Electric Kettle

🟧 I want to make some hot tea, what can I use to boil water?

Ice Bucket

🟩 I have some drinks that need to stay cold; where can I put some ice cubes?

Luggage Rack

🟩 I want to unpack my bag without putting it on the floor or the bed; where should I place it?

Suitcase

🟥 I need to pack my clothes to leave the hotel, what should I use?

Television

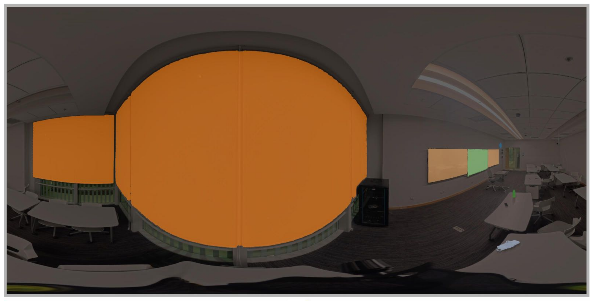
🟥 I'd like to catch up on the news; which screen should I turn on?

Water Bottle

🟪 I'm feeling very thirsty after my flight, which object can I use to get a drink of water?

Figure 16 Visualizations of PAP-12K(Bedroom).

Classroom Case 13



Clock

🟦 Where can I find a device on the wall to check if I am running late for my next appointment?

Power Strip

🟦 I need to charge my phone and laptop at the same time; which object on the desk provides multiple electrical outlets?

Roller Blinds

🟧 The sunlight is reflecting off the screen and making it hard to see; what can I use to cover the windows?

Television

🟧 Which object allows me to display the presentation from my laptop to all the students in the room?

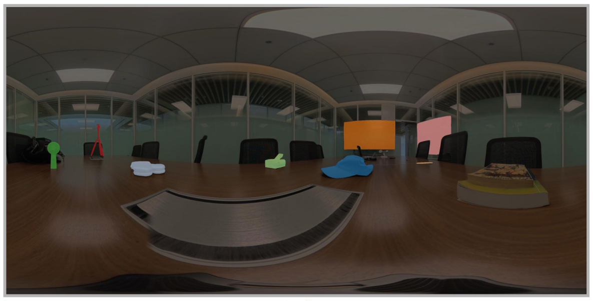
Water Bottle

🟩 I am feeling dehydrated; which object on the desk contains something for me to drink?

Whiteboard

🟩 Which surface allows me to illustrate a concept to my colleagues with a marker?

Classroom Case 14



Hat

🟦 Which item can I put on my head to keep it warm or style my outfit?

Headphones

🟦 I want to join a call without the audio playing through the room's speakers; what can I use?

Monitor

🟧 Which large electronic display at the end of the table can be used to present a slideshow?

Pen

🟧 I need to write a quick note on a piece of paper; what tool can I find on the table to do so?

Portable Fan

🟩 The room feels a bit warm; what small white device can I use to blow cool air towards me?

Tissue Box

🟩 I need to sneeze; where can I find a box containing soft paper tissues?

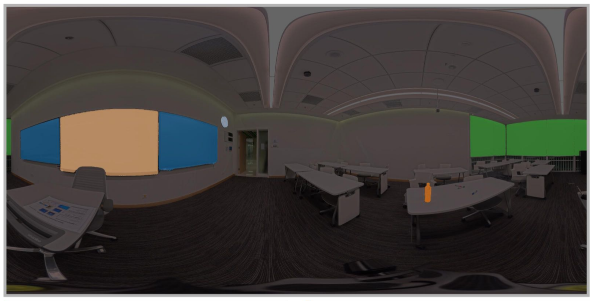
Tripod

🟥 I need a stable mount for my smartphone to take a group photo; what should I use?

Whiteboard

🟥 Where can I use a marker to draw a diagram that everyone in the room can see?

Classroom Case 15



Large Screen

🟦 I need to display my laptop's screen for everyone to see; which large wall-mounted device should I use?

Wall Clock

🟦 I want to check if the class is almost over; where can I see the time?

Water Bottle

🟧 I am thirsty and need a drink; which object on the table contains a beverage?

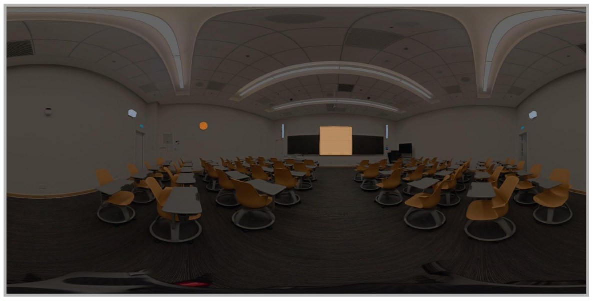
Whiteboard

🟧 I want to use a dry-erase marker to brainstorm some ideas; which surface is designed for this?

Window Blinds

🟩 The sun's reflection is too bright on the table; what can I use to cover the windows?

Classroom Case 16



Exit Sign

🟦 If I need to leave the room quickly, which illuminated green sign above the door should I follow?

Speaker

🟦 I want to play some audio for the class; which wall-mounted device will project the sound throughout the room?

Wall Clock

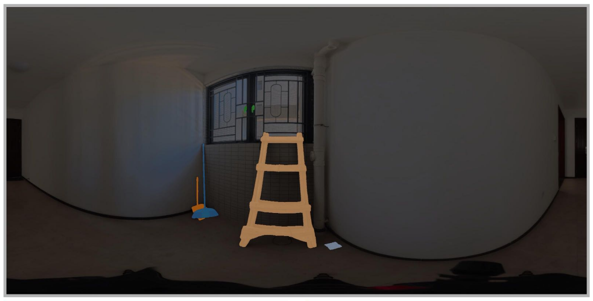
🟧 I'm worried about being late for my next meeting; where can I look to check the current time?

Whiteboard

🟧 The teacher needs a large white surface in the center of the front wall to write with markers; where is it?

Figure 17 Visualizations of PAP-12K(Classroom).

Corridor Case 17



Broom

🟦 There is dust and debris on the floor, what can I use to gather it together?

Drain

🟦 If I spill water on the floor, where is the designated spot for it to flow out of the room?

Dustpan

🟧 After sweeping the dirt into a pile, what can be used to collect the trash?

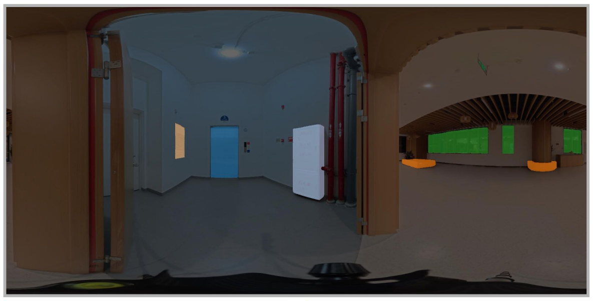
Ladder

🟧 I need to reach the top of the window frame to clean it, what can I climb on?

Window Handle

🟩 I want to let some fresh air into this room, what should I operate?

Corridor Case 18



Elevator

🟦 I need to travel to a different floor; which object should I approach?

Fire Hydrant Cabinet

🟦 Where can I find a cabinet that contains emergency fire-fighting gear?

Rectangular Wooden Platform

🟧 I want to take a break and sit down; where is the low, rectangular wooden structure located in the open room?

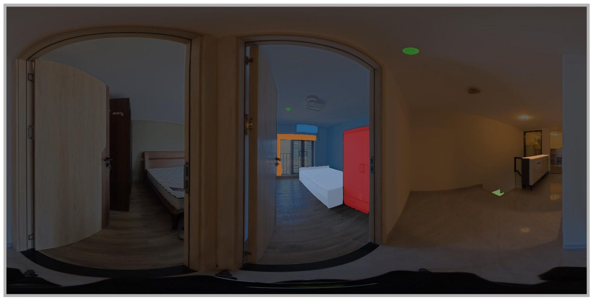
Vent

🟧 Where is the wall-mounted grate used for airflow in the hallway?

Windows

🟩 I want to see the view of the outdoors; where should I look?

Corridor Case 19



Air Conditioner

🟦 The bedroom is feeling very stuffy and hot; what wall-mounted device can I use to lower the temperature?

Bed

🟦 I'm exhausted and need a place to sleep in the room with the blue walls; where should I go?

Curtains

🟧 The light from the balcony is too bright for my nap; what can I pull across to darken the room?

Sideboard

🟧 I'm looking for a long, flat surface in the hallway to display some decorative items; what can I use?

Smoke Detector

🟩 Which safety device on the ceiling is designed to detect a fire and sound an alarm?

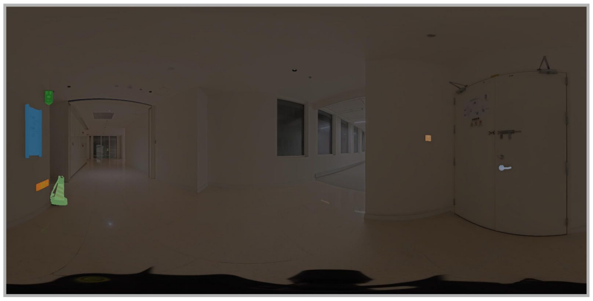
Stairs

🟩 I want to go down to the next level of the house; what should I walk down?

Wardrobe

🟥 I have many clothes to hang up in the blue room; which piece of furniture has doors for storage?

Corridor Case 20



Directory Sign

🟦 I am looking for room 512; which wall-mounted board provides a list of room ranges and their directions?

Door Handle

🟦 What part of the door should I grip and turn to unlatch it?

Exit Sign

🟧 If I need to leave the building quickly in an emergency, which glowing green object should I follow?

Keypad

🟧 The door is electronically locked; where can I swipe my badge to get inside?

Restroom Sign

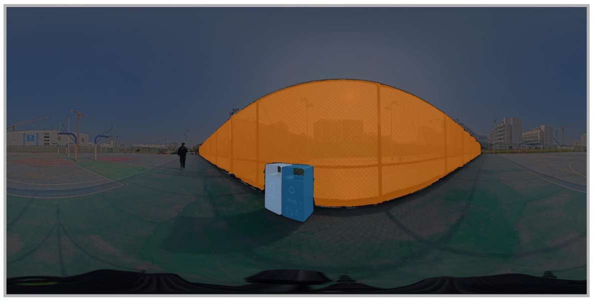
🟩 I am looking for a place to wash my hands; which blue sign hanging from the wall will guide me?

Traffic Cone

🟩 There is a hazard on the floor; what object can I place near it to warn people to walk around?

Figure 18 Visualizations of PAP-12K(Corridor).

Gym Case 21



Non-Recyclable Trash Can

🟦 I need to dispose of some non-recyclable waste; which bin is intended for general trash?

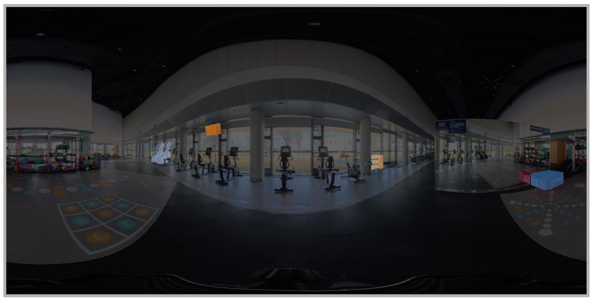
Recyclable Trash Can

🟦 I have an empty soda can that I want to recycle; where should I put it?

Chain-Link Fence

🟧 What structure is used to prevent balls from flying out of the court area?

Gym Case 22



Plyometric Box

🟦 I want to practice my explosive jumping power; what large red padded block can I jump onto?

Stair Climber

🟦 I want to simulate walking up an endless flight of stairs; which machine is designed for this?

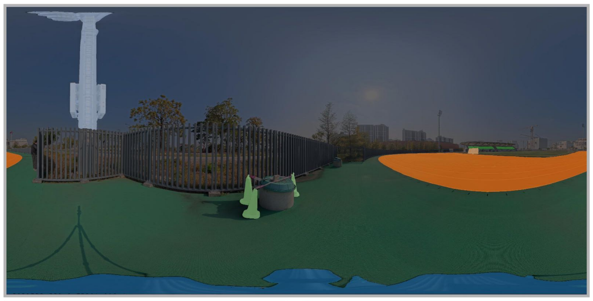
Television

🟧 I want to watch some media while I use the cardio machines; where is the screen located?

Water Jugs

🟧 I am very thirsty after my workout; where can I get some water to drink?

Gym Case 23



Camera Mount

🟦 I need a stable base to hold my camera perfectly still for a panoramic shot; what should I use?

Floodlight Pole

🟦 Which structure allows the sports field to be used for games after the sun goes down?

Running Track

🟧 I want to practice my 100-meter sprint, which part of the ground should I use?

Soccer Goal

🟧 I want to play a game of football; where should I kick the ball to score a point?

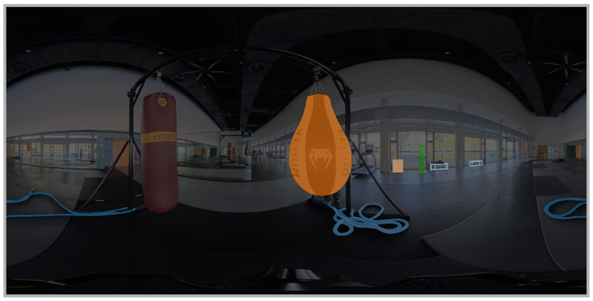
Stadium Seating

🟩 Where can I find a place for hundreds of people to sit and watch the race?

Traffic Cone

🟩 There is a trip hazard near the fence; what can I place there to warn others?

Gym Case 24



Battle Rope

🟦 I want to perform high-intensity waves and slams using the thick, coiled ropes on the floor.

Bench

🟦 I am feeling tired after my workout and need a simple place to sit down and rest.

Teardrop Punching Bag

🟧 I need the white and black teardrop-shaped bag to practice my uppercuts and timing.

Trash Can

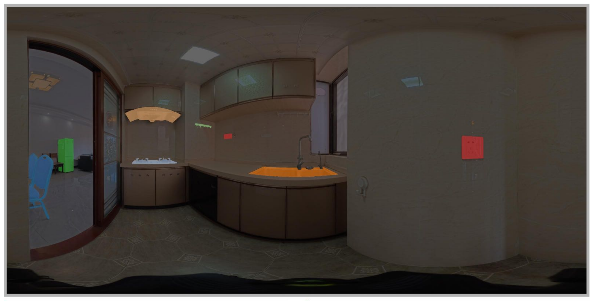
🟧 Where can I dispose of my empty water bottle or other waste in this room?

Weight Scale

🟩 Which device near the windows can I use to measure my body weight and height?

Figure 19 Visualizations of PAP-12K(Gym).

Kitchen Case 25



Dining Chair

■ I want to sit down while I wait for the food to cook; which red object in the background is available?

Gas Stove

■ I need to cook some pasta; what can I use to heat the pot?

Kitchen Sink

■ Where can I find a place to wash my hands after handling raw meat?

Range Hood

■ The kitchen is getting too hot and smoky; what can I use to suck the air out?

Refrigerator

■ Where can I find a cold storage area for my groceries?

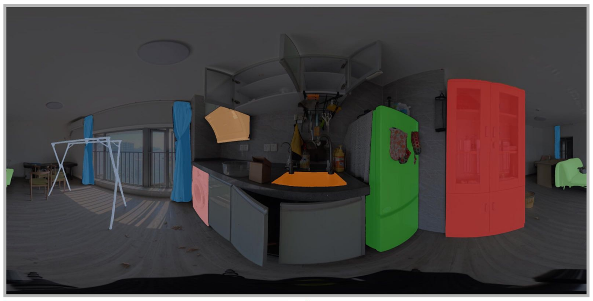
Wall Hook

■ Where is a convenient place to hang a damp cloth near the sink?

Wall Outlet

■ How can I power my electric kettle using the interface on the tiled wall?

Kitchen Case 26



Curtains

■ It's getting dark and I want to block the view from the street; what can I draw across the large windows?

Drying Rack

■ I've just hand-washed a delicate shirt; where is a suitable place to hang it so it can dry in the sun?

Kitchen Sink

■ I need to fill a pot with water to boil pasta; where is the most convenient place to do this?

Range Hood

■ I am searing a steak and there is a lot of smoke; what should I activate to clear the air?

Refrigerator

■ Where can I find a chilled environment to keep my leftovers fresh for tomorrow?

Sofa

■ I want to relax and read a book in a soft, cushioned seat; where should I go?

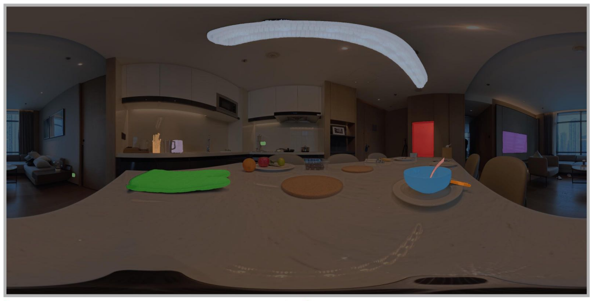
Tall Wooden Cabinet

■ I have some extra glasses and plates that don't fit in the kitchen; where can I store them?

Washing Machine

■ My gym clothes are sweaty and need a deep clean; where can I put them to be laundered?

Kitchen Case 27



Bowl

■ I want to have some hot soup for lunch, which container on the marble table should I use?

Chandelier

■ The room is getting dark; what large, curved crystal fixture on the ceiling can I turn on for light?

Chopsticks

■ I am eating noodles and prefer using traditional East Asian utensils; what should I pick up from the table?

Knife Block

■ I need to slice some vegetables; where can I find the set of sharp tools kept in a wooden holder?

Oven Mitt

■ I need to take a hot tray out of the oven; what on the island can I wear to protect my hand from burns?

Power Outlet

■ My phone is running low on battery; where on the surface of the kitchen island can I plug in my charger?

Refrigerator

■ I want a cold glass of water; which tall appliance in the kitchen is used for chilling drinks?

Spoon

■ I need a utensil to eat my soup without making a mess; what should I use?

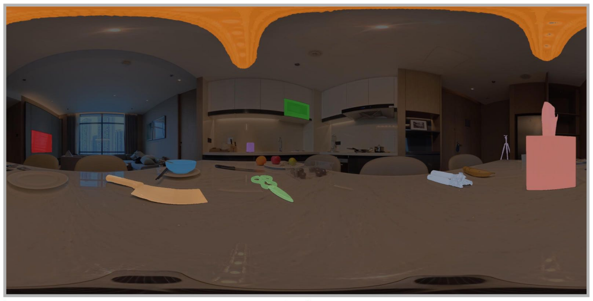
Television

■ I want to keep up with the latest news on a large screen; what device in the living area should I turn on?

Toaster

■ I want my bread to be crispy and warm for breakfast; which silver appliance on the back counter should I use?

Kitchen Case 28



Bowl

■ I want to eat some cereal with milk; what container on the table is best for holding liquid food?

Can Opener

■ I want to eat some canned soup, but the lid doesn't have a pull tab; what device on the counter can help me open it?

Chandelier

■ It's getting dark outside; what can I turn on to provide bright, decorative lighting for the whole kitchen area?

Cleaver

■ I have a large piece of meat with bone that needs to be chopped; what tool on the kitchen island is best for this?

Microwave

■ I have some leftover pasta that is cold; which built-in appliance can I use to heat it up in just a minute?

Scissors

■ I need to open a sealed plastic bag of frozen vegetables; what can I use on the counter to cut it?

Television

■ Where can I watch a movie or the evening news while sitting on the sofa in the background?

Tissue Box

■ I accidentally spilled a small amount of water on the counter and need something disposable to wipe it up.

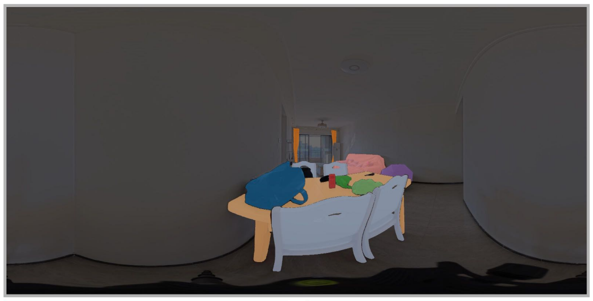
Toaster

■ I want my bread to be crispy and brown for breakfast; which small appliance on the back counter should I use?

Tripod

■ I want to take a steady long-exposure photo of the city view; what equipment in the corner by the door can hold my camera?

Livingroom Case 29



Bag

🟦 I need to carry my books and personal items to the library; which container on the table is suitable?

Chair

🟦 I'm tired of standing while eating; what can I use to sit down at the table?

Curtains

🟧 The sunlight coming through the balcony door is too bright; what can I use to dim the room?

Dining Table

🟧 Where can I find a large, flat surface to spread out my work papers and laptop?

Hairdryer

🟩 My hair is wet after a shower; what device on the table can I use to dry it?

Hat

🟩 I want to protect my head from the sun while I'm out; what accessory on the table can I wear?

Remote Control

🟥 I want to adjust the temperature of the air conditioner; what handheld device on the table should I use?

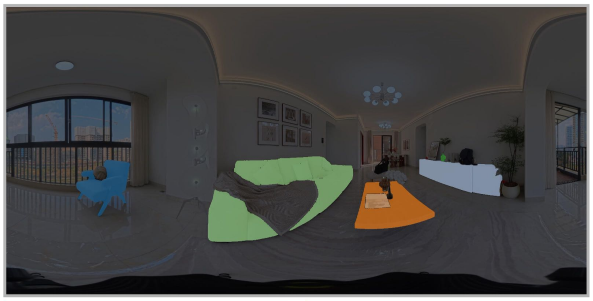
Sofa

🟥 I'm looking for a comfortable, cushioned place to lounge and relax in the background.

Umbrella

🟪 It's raining outside; what object leaning near the sofa can I take to stay dry?

Livingroom Case 30



Armchair

🟦 I want a private spot by the window to sit and enjoy the view; which single-person seat is best?

Cabinet

🟦 I need a long, low surface to store my bags and display a potted plant; what piece of furniture should I use?

Coffee Table

🟧 I'm finished with my drink; where is a flat surface near the sofa to set my cup down?

Magazines

🟧 I'm bored and want to read; where can I find some printed media on the coffee table?

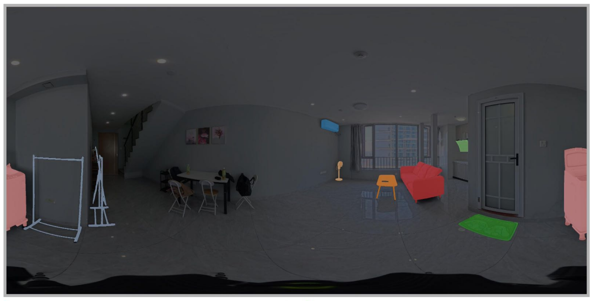
Potted Plant

🟩 I want to add some nature to the room; which object provides greenery next to the white cabinet?

Sofa

🟩 I have several guests over; where can we all sit together comfortably in this room?

Livingroom Case 31



Air Conditioner

🟦 The summer heat is making the room uncomfortable; what can I use to lower the temperature?

Clothes Rack

🟦 My clothes are wet after being washed; where can I hang them to dry?

Coffee Table

🟧 I have a cup of coffee and I'm sitting on the couch; where can I safely set my mug down?

Electric Fan

🟧 I want a gentle breeze while I'm sitting; which portable appliance can I turn on?

Floor Mat

🟩 My shoes are a bit dusty from outside; where can I wipe them before walking further into the room?

Kitchen Hood

🟩 I'm frying food and it's getting smoky; which device can I use to clear the air?

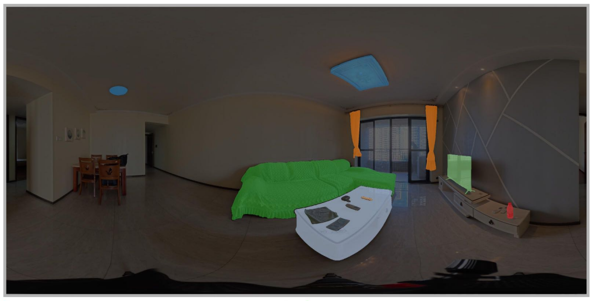
Sofa

🟥 I'm feeling tired and want to lie down on something soft; where is the best place in the living room?

Washing Machine

🟥 I have a pile of dirty clothes; which appliance can I use to wash them?

Livingroom Case 32



Ceiling Light

🟦 It's getting dark outside; what can I switch on to brighten up the entire living area?

Coffee Table

🟦 I have a cup of tea and want to set it down while I sit on the couch; which surface is most convenient?

Curtains

🟧 The glare on the TV is too strong; what can I use to block the light coming from the balcony?

Remote Control

🟧 I want to turn up the volume on the TV from the sofa; what handheld device on the coffee table do I need?

Sofa

🟩 I'm feeling very tired and want to take a nap; where is the best place to lie down in this room?

Television

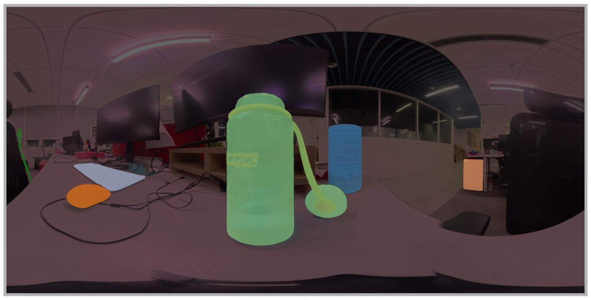
🟩 Which electronic device in the room can I use to catch up on the latest news broadcasts?

Water Bottle

🟥 I'm feeling thirsty; is there a container of liquid I can drink from on the TV stand?

Figure 21 Visualizations of PAP-12K(Livingroom).

Office Case 33



Collagen Peptides Container

🟦 I want to prepare a chocolate-flavored nutritional supplement; which white tub should I open?

Keyboard

🟦 I need to type an email to my colleague; what device should I use to enter the text?

Mouse

🟧 I need to click on an icon on the screen; which pink object should I move with my hand?

Red Cabinet

🟧 I have some sensitive documents to put away; which red storage unit with drawers can I use?

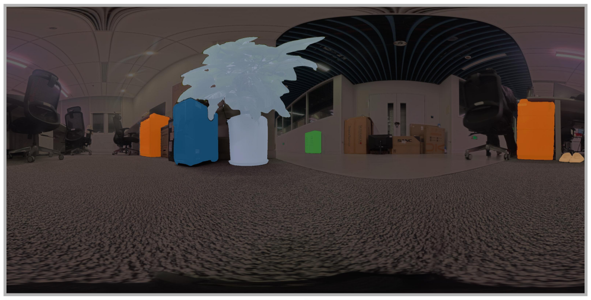
Tripod

🟩 I want to take a steady photograph; what can I use to mount and stabilize my camera?

Water Bottle

🟩 I am feeling thirsty; which blue container on the desk can I use to take a drink of water?

Office Case 34



Computer Tower

🟦 I need to power on a desktop computer to process some data; which large black device should I press the button on?

Plant

🟦 The room feels too sterile; which object adds a touch of nature and greenery to the space?

Red Cabinet

🟧 I have some private documents I need to lock away; where can I store them?

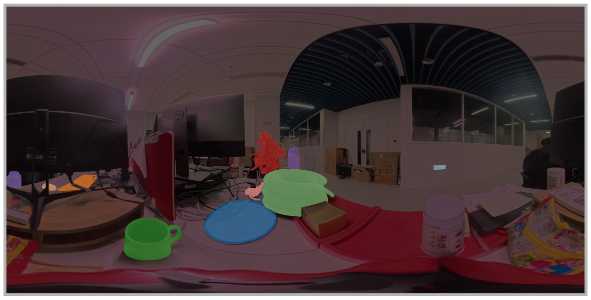
Slippers

🟧 My feet are tired from wearing dress shoes; what comfortable footwear can I change into under the desk?

Trash Can

🟩 I have some scrap paper to throw away; where is the designated place for waste?

Office Case 35



Badminton Racket

🟦 I want to go to the gym for a game of badminton; what equipment on the desk do I need?

Exit Sign

🟦 In case of an emergency, which sign should I follow to find the nearest way out?

Keyboard

🟧 How can I type out an email to my supervisor using the equipment on this desk?

Mouse

🟧 I need to click on a link on the screen; which small device should I move with my hand?

Mug

🟩 I'm thirsty and want to pour some hot tea; what container on the desk is suitable?

Neck Pillow

🟩 My neck is getting stiff from sitting too long; what can I use for support while I take a quick nap?

Potted Plant

🟥 The office feels a bit sterile; what can I look at to see some natural greenery?

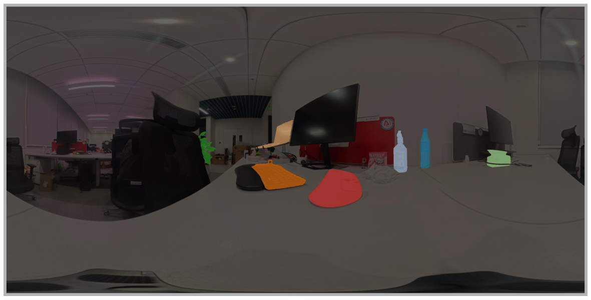
Power Strip

🟥 I have multiple devices to plug in but only one wall socket; what can I use to expand the number of outlets?

Water Bottle

🟪 I need to stay hydrated throughout the day; what portable container can I fill with water?

Office Case 36



Green Bottle

🟦 I am thirsty and looking for a drink; what contains a liquid I can consume?

Hand Sanitizer

🟦 I want to clean my hands before eating; which bottle contains disinfectant?

Keyboard

🟧 I need to type a report; what should I use to enter the characters?

Laptop

🟧 I need to go to a meeting; which computer can I easily pick up and take with me?

Potted Plant

🟩 I want to see something natural and green in this office; where should I look?

Tissues

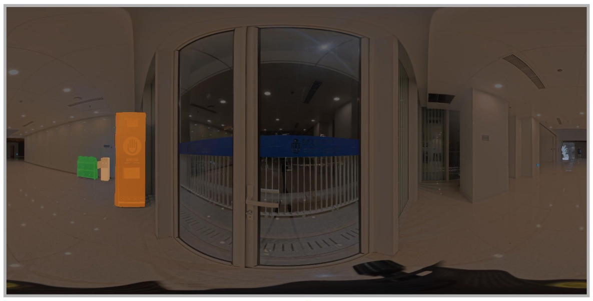
🟩 I need to wipe my face; where can I find some soft paper?

Vertical Mouse

🟥 My wrist is hurting; which ergonomic peripheral on the desk can I use to control the cursor?

Figure 22 Visualizations of PAP-12K(Office).

Pantry Case 37



Exit Sign

🟦 I need to identify the direction of the nearest exit; which small green illuminated sign should I look at?

Potted Plant

🟦 I'm looking for some greenery to brighten up the white corridor; where is the indoor plant located?

Vending Machine

🟧 I am looking for a place to buy a quick snack or drink; which blue machine should I go to?

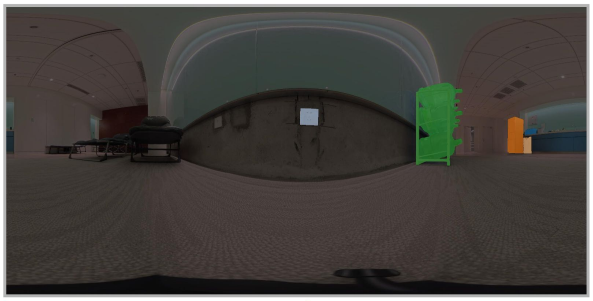
Water Dispenser

🟧 I need to get a glass of water; which device in the corner is designed to pour it for me?

Water Jugs

🟩 Where are the large blue plastic containers that hold the reserve drinking water for the room?

Pantry Case 38



Microwave

🟦 My dinner has gone cold; what can I use to heat it up quickly?

Power Outlet

🟦 My laptop is about to shut down due to low battery; what can I use to charge it?

Refrigerator

🟧 Where can I find a cold place to store my perishable groceries?

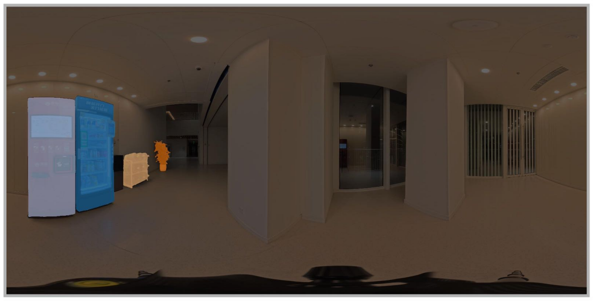
Trash Can

🟧 Where can I find a designated place to throw away my empty food wrappers?

Water Jugs

🟩 Where can I find a large supply of clean water to stay hydrated?

Pantry Case 39



Beverage Vending Machine

🟦 I am thirsty for a cold soda or bottled tea; where should I go?

Coffee Vending Machine

🟦 I'm feeling sleepy and need a hot caffeine boost; which machine should I use?

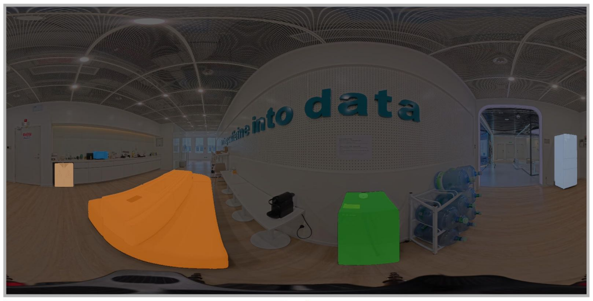
Potted Plant

🟧 I want to add a touch of nature to this sterile environment; which object provides some greenery?

Water Bottle Rack

🟧 The office water cooler is empty; where can I find a replacement 5-gallon jug?

Pantry Case 40



Microwave

🟦 My soup has gone cold; which device on the counter can I use to heat it up quickly?

Refrigerator

🟦 My lunch needs to stay chilled until noon; where is the best place to store it?

Sofa

🟧 I've been standing all day and need a comfortable place to sit and rest; where should I go?

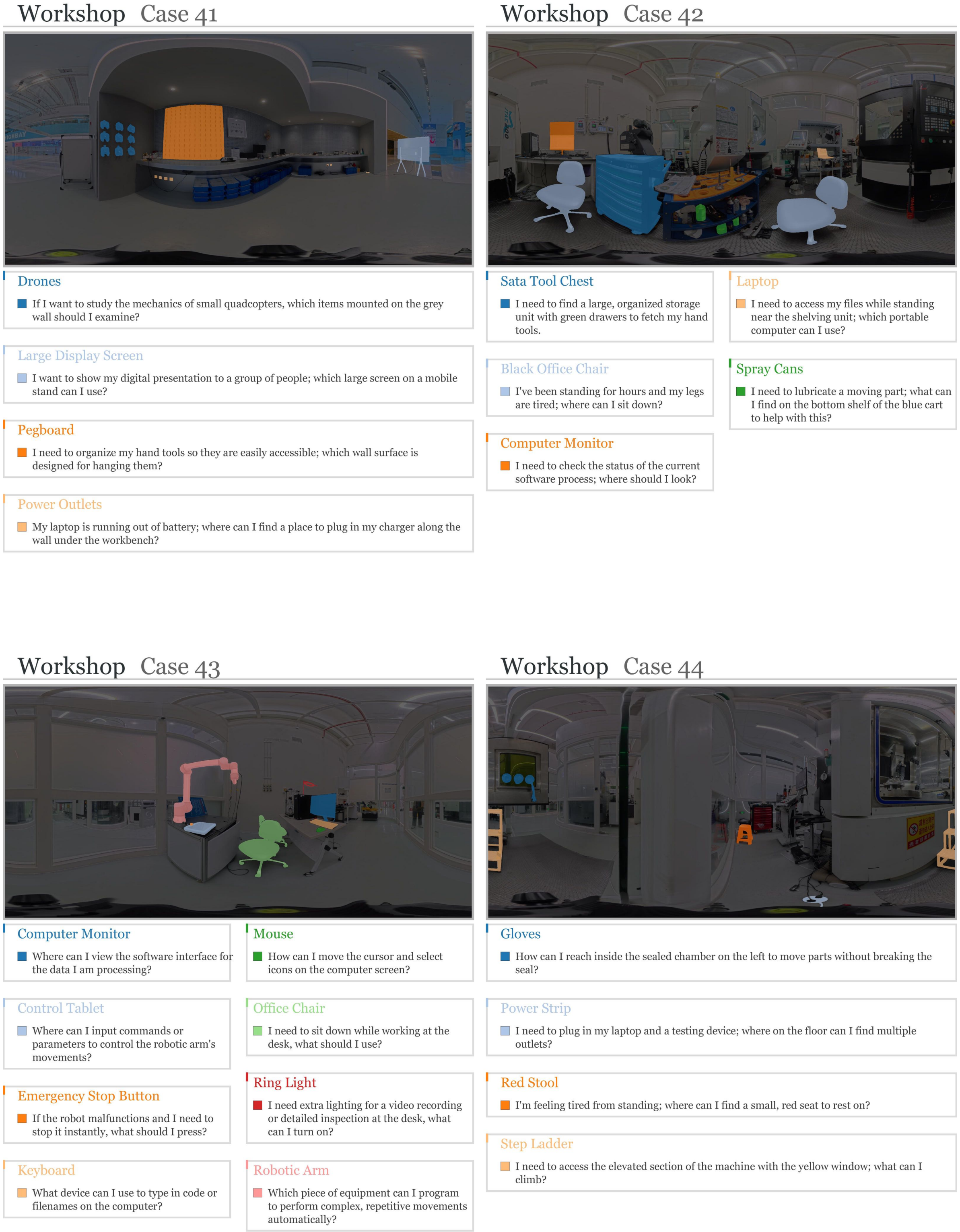
Trash Cans

🟧 I have an empty plastic bottle; where can I dispose of it for recycling?

Water Dispenser

🟩 I'm feeling dehydrated; what black machine can provide me with a glass of drinking water?

Figure 23 Visualizations of PAP-12K(Pantry).



Workshop Case 41

Drones

🟦 If I want to study the mechanics of small quadcopters, which items mounted on the grey wall should I examine?

Large Display Screen

🟦 I want to show my digital presentation to a group of people; which large screen on a mobile stand can I use?

Pegboard

🟧 I need to organize my hand tools so they are easily accessible; which wall surface is designed for hanging them?

Power Outlets

🟧 My laptop is running out of battery; where can I find a place to plug in my charger along the wall under the workbench?

Workshop Case 42

Sata Tool Chest

🟦 I need to find a large, organized storage unit with green drawers to fetch my hand tools.

Black Office Chair

🟦 I've been standing for hours and my legs are tired; where can I sit down?

Computer Monitor

🟧 I need to check the status of the current software process; where should I look?

Laptop

🟧 I need to access my files while standing near the shelving unit; which portable computer can I use?

Spray Cans

🟩 I need to lubricate a moving part; what can I find on the bottom shelf of the blue cart to help with this?

Workshop Case 43

Computer Monitor

🟦 Where can I view the software interface for the data I am processing?

Control Tablet

🟦 Where can I input commands or parameters to control the robotic arm's movements?

Emergency Stop Button

🟧 If the robot malfunctions and I need to stop it instantly, what should I press?

Keyboard

🟧 What device can I use to type in code or filenames on the computer?

Mouse

🟩 How can I move the cursor and select icons on the computer screen?

Office Chair

🟩 I need to sit down while working at the desk, what should I use?

Ring Light

🟥 I need extra lighting for a video recording or detailed inspection at the desk, what can I turn on?

Robotic Arm

🟥 Which piece of equipment can I program to perform complex, repetitive movements automatically?

Workshop Case 44

Gloves

🟦 How can I reach inside the sealed chamber on the left to move parts without breaking the seal?

Power Strip

🟦 I need to plug in my laptop and a testing device; where on the floor can I find multiple outlets?

Red Stool

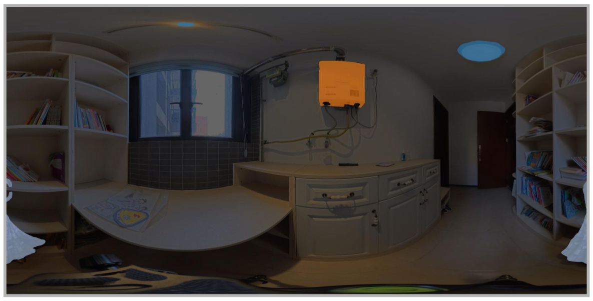
🟧 I'm feeling tired from standing; where can I find a small, red seat to rest on?

Step Ladder

🟧 I need to access the elevated section of the machine with the yellow window; what can I climb?

Figure 24 Visualizations of PAP-12K(Workshop).

Others Case 45



Ceiling Light

🟦 Which fixture can I turn on to provide general illumination from above when the room is too dark?

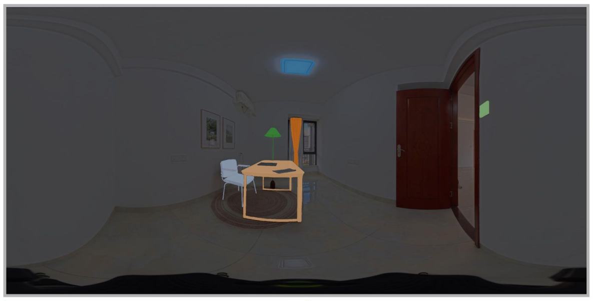
Umbrella

🟦 I am about to head out into a rainstorm; which object leaning against the shelf can I use to stay dry?

Water Heater

🟧 Which wall-mounted appliance is used to provide hot water for the household's needs?

Others Case 46



Ceiling Light

🟦 I just entered a dark room; what should I turn on to illuminate the entire space from above?

Chair

🟦 I am tired of standing and need a place to sit while I use the desk.

Curtains

🟧 The sunlight coming through the window is too bright; what can I use to shade the room?

Desk

🟧 I need a flat, stable surface to spread out my documents and write; where should I work?

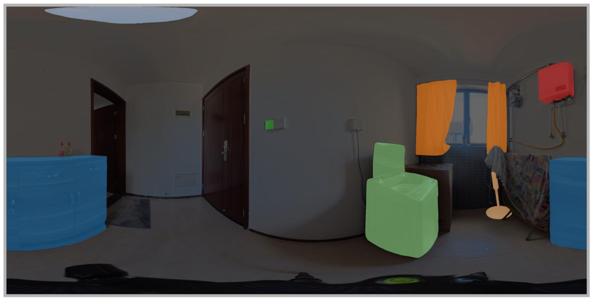
Floor Lamp

🟩 The room is a bit dim for reading; what can I use to provide extra light specifically over the desk area?

Light Switch

🟩 I want to adjust the lighting; what should I toggle on the wall near the wooden door?

Others Case 47



Cabinet

🟦 Where can I find a piece of furniture with drawers to organize and store my small belongings?

Ceiling Light

🟦 Which fixture is responsible for providing light from above to the entire room?

Curtains

🟧 The glare from the window is making it hard to see; what can I pull across to block the light?

Electric Fan

🟧 The room feels stuffy and hot; what can I uncover and turn on to create a breeze?

Light Switch

🟩 It is getting dark; what can I press on the wall near the door to turn on the overhead light?

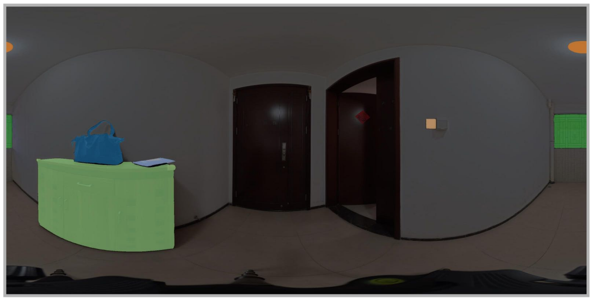
Washing Machine

🟩 My clothes are dirty after a long hike; which appliance can I use to get them clean again?

Water Heater

🟥 I want to take a warm shower; which device on the wall is responsible for heating the water?

Others Case 48



Black Duffel Bag

🟦 I'm going on a weekend trip; which object can I use to pack my clothes and toiletries?

Blue Envelope

🟦 I received a small package in the mail; where is it located in this room?

Ceiling Light

🟧 Which object is the primary source of artificial light in this space?

Light Switch

🟧 It's getting late and I can't see well; what should I toggle on the wall to brighten the room?

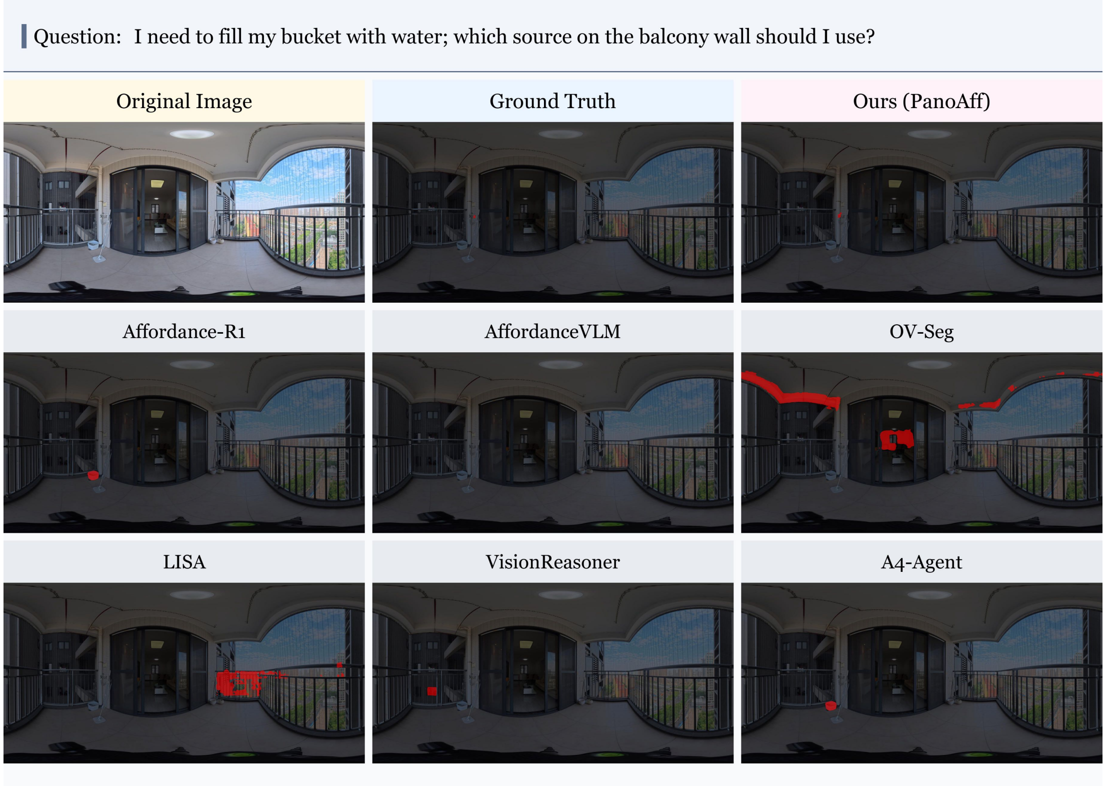
Window

🟩 I want to see what the weather is like outside; where should I look?

Wooden Cabinet

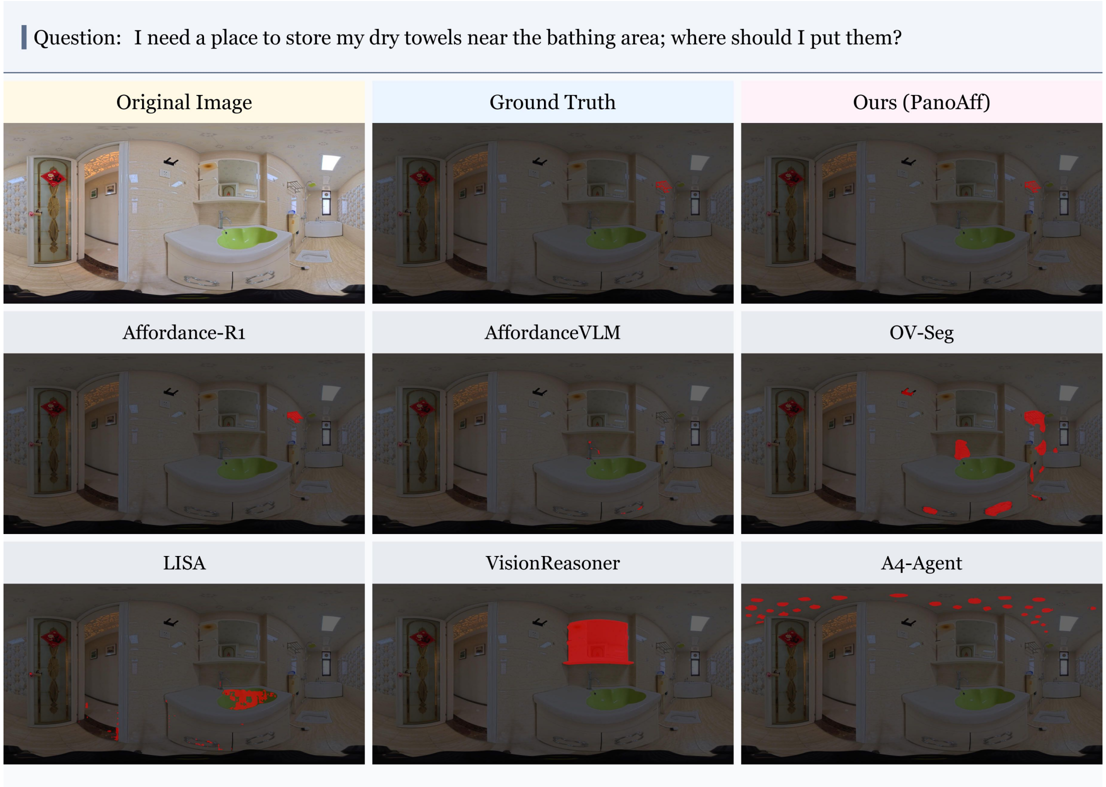
🟩 I just got home and want to set down my heavy bag; where is a suitable waist-high surface to place it?

Figure 25 Visualizations of PAP-12K(Others).



Question: I need to fill my bucket with water; which source on the balcony wall should I use?

Figure 26 Qualitative comparison on PAP-12K (Balcony).



Question: I need a place to store my dry towels near the bathing area; where should I put them?

Figure 27 Qualitative comparison on PAP-12K (Bathroom).

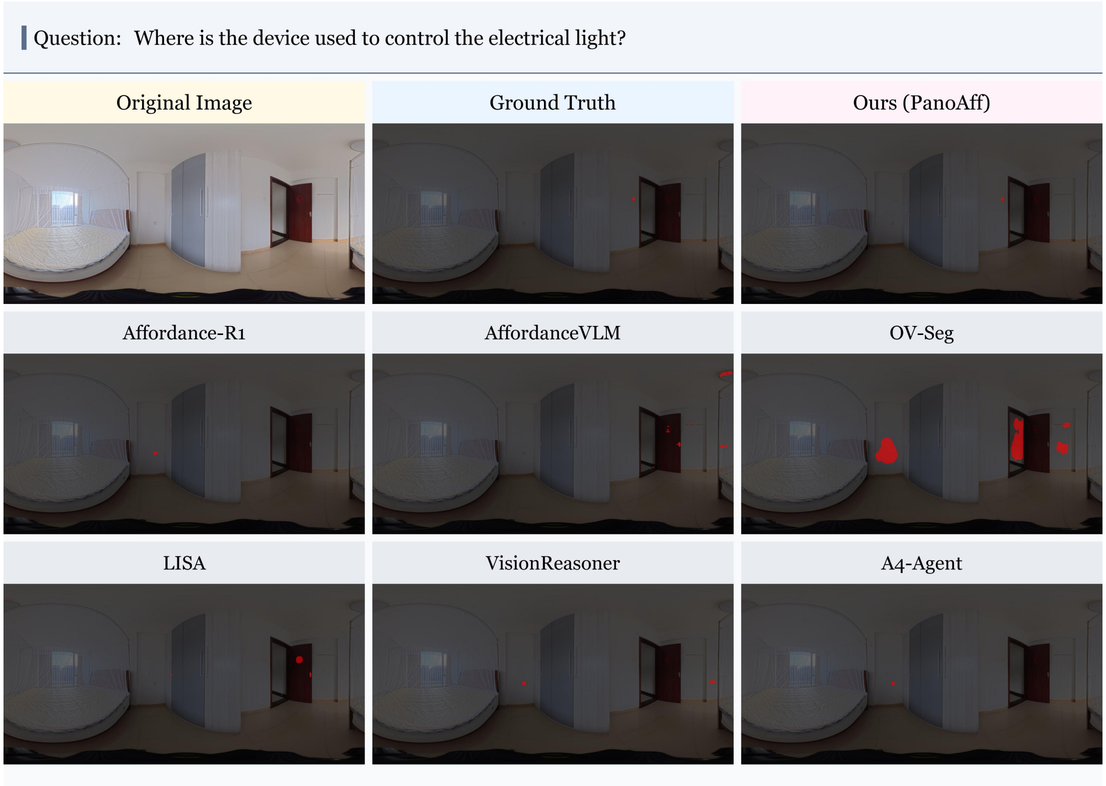


Figure 28 Qualitative comparison on PAP-12K (Bedroom).

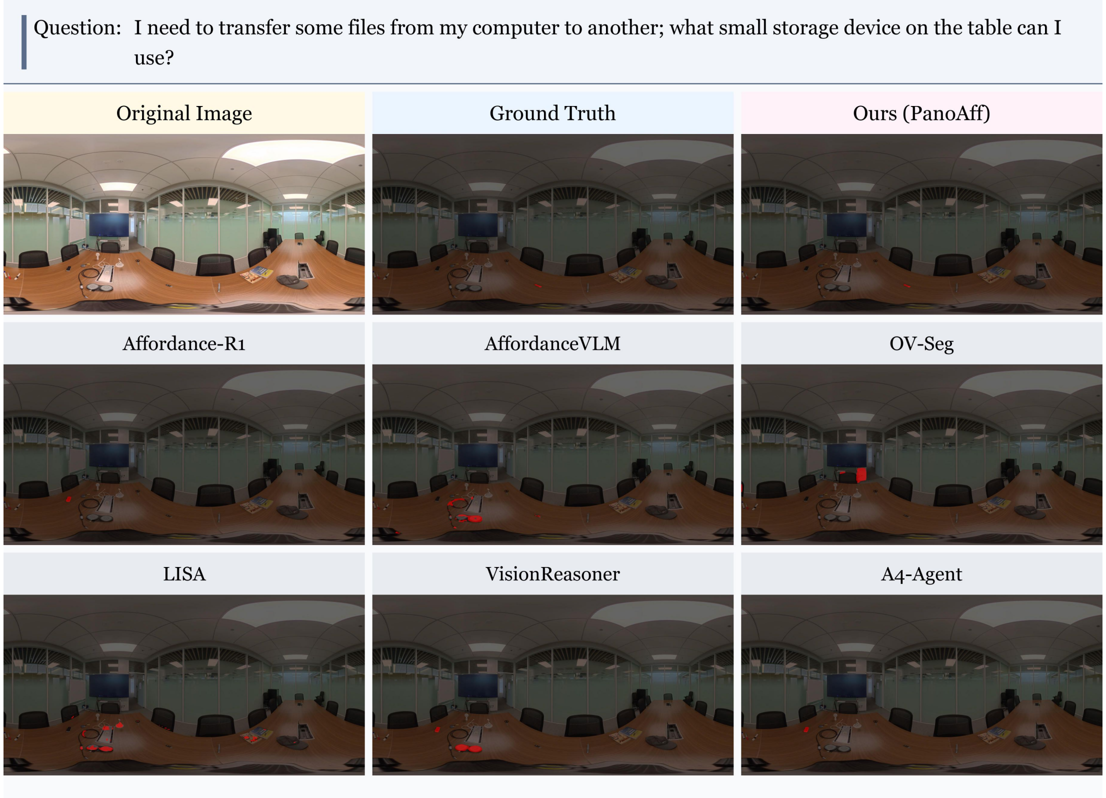
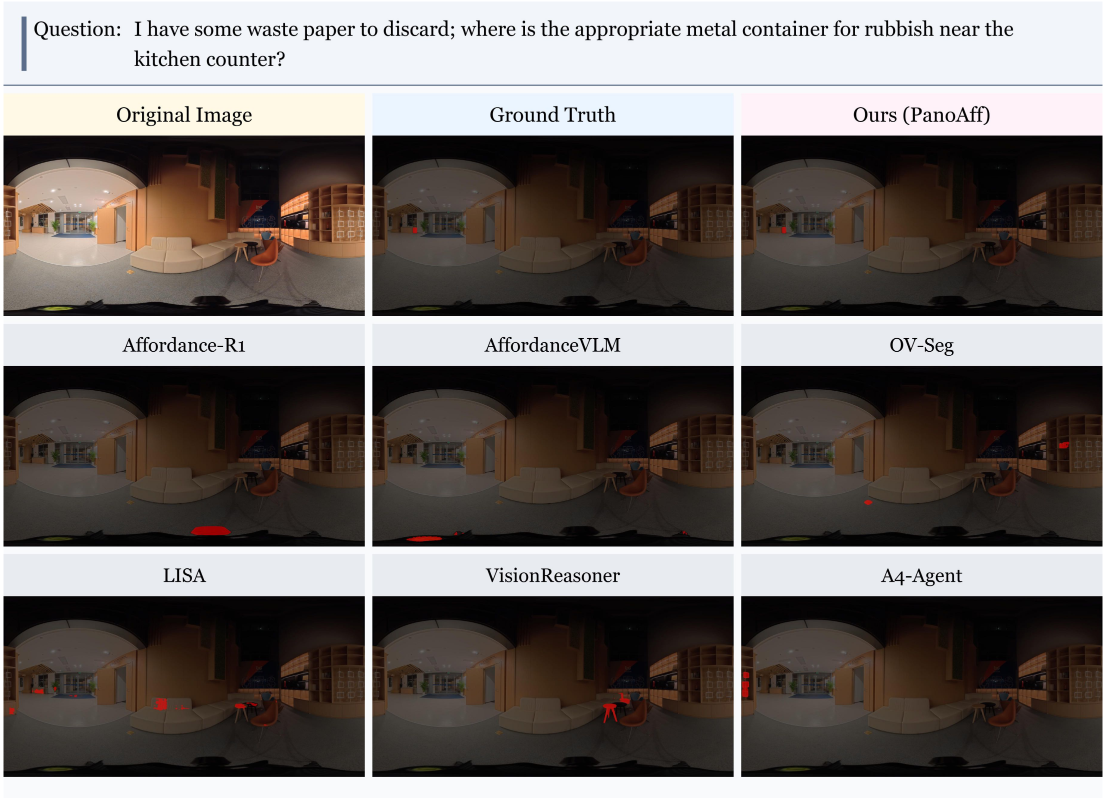
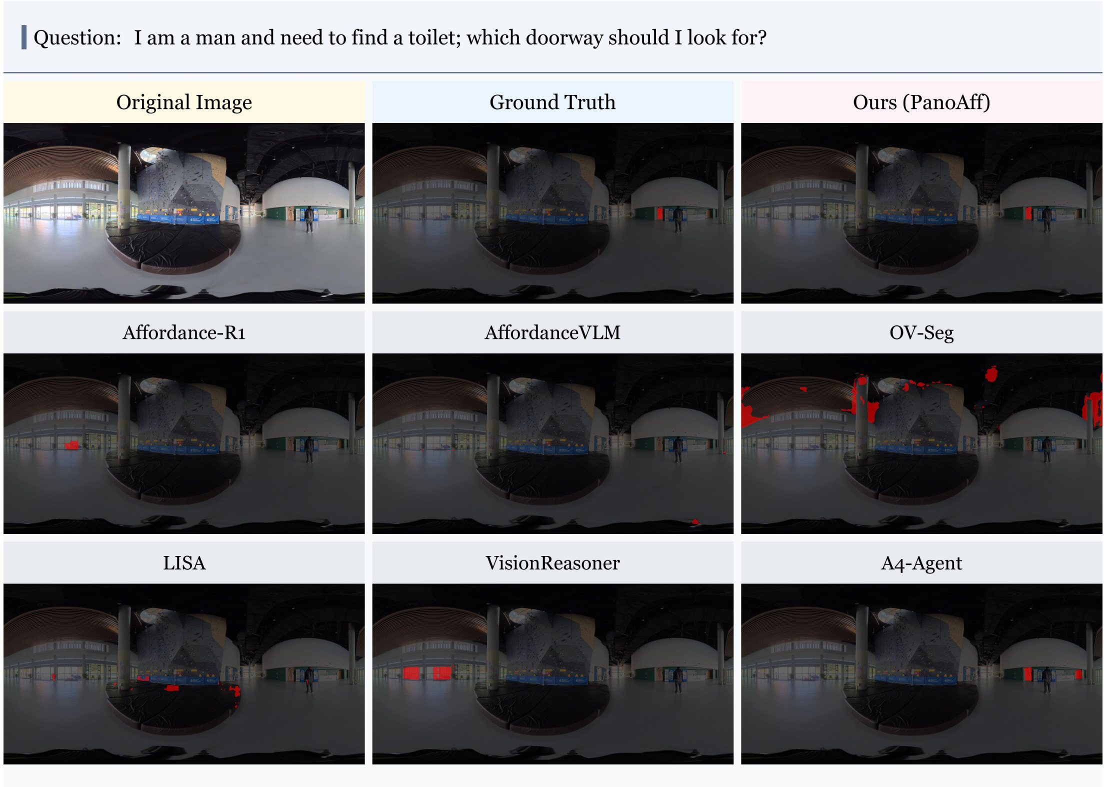

Question: I need to transfer some files from my computer to another; what small storage device on the table can I use?
Original Image
Ground Truth
Ours (PanoAff)
Affordance-R1
AffordanceVLM
OV-Seg
LISA
VisionReasoner
A4-Agent

Figure 29 Qualitative comparison on PAP-12K (Classroom).



Question: I have some waste paper to discard; where is the appropriate metal container for rubbish near the kitchen counter?

Figure 30 Qualitative comparison on PAP-12K (Corridor).



Question: I am a man and need to find a toilet; which doorway should I look for?

Figure 31 Qualitative comparison on PAP-12K (Gym).

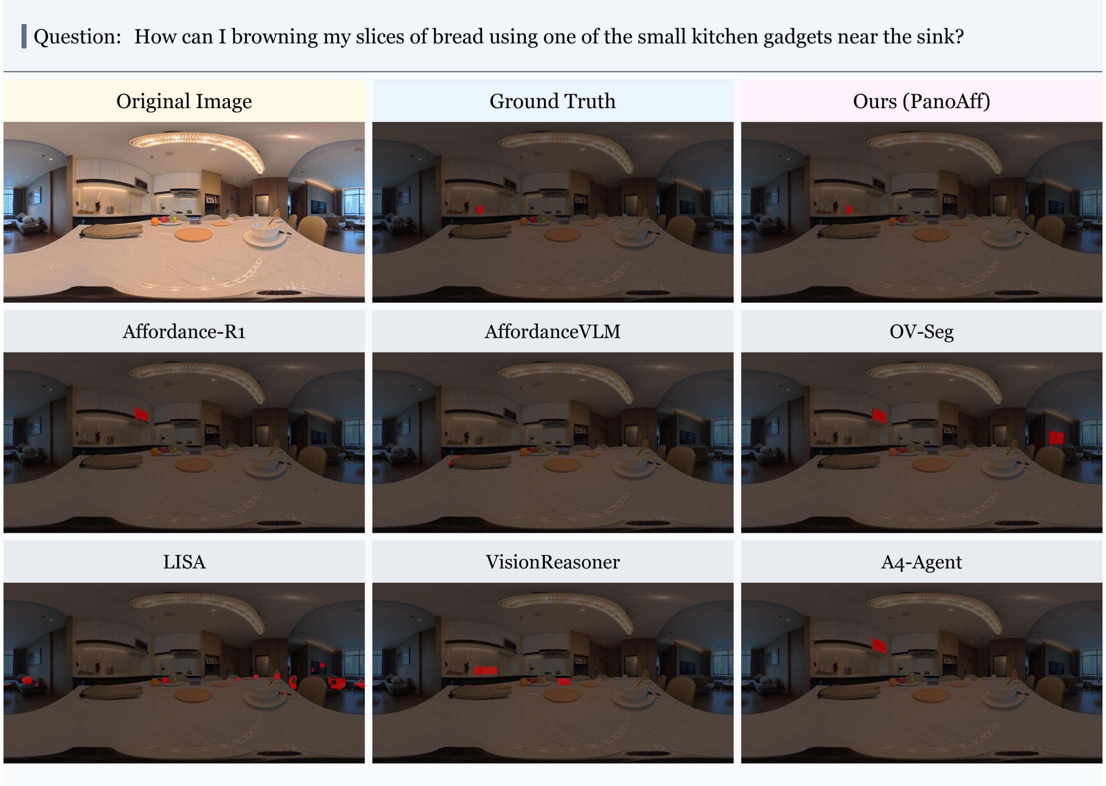

Question: How can I browning my slices of bread using one of the small kitchen gadgets near the sink?

Figure 32 Qualitative comparison on PAP-12K (Kitchen).

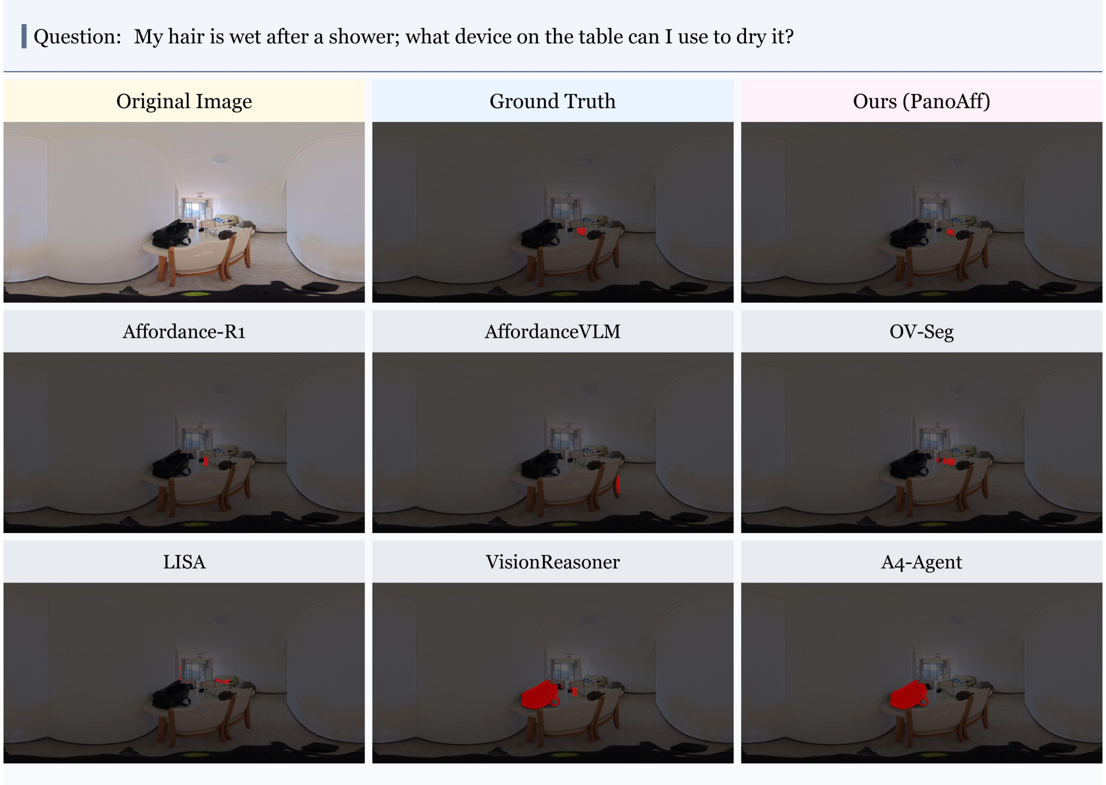
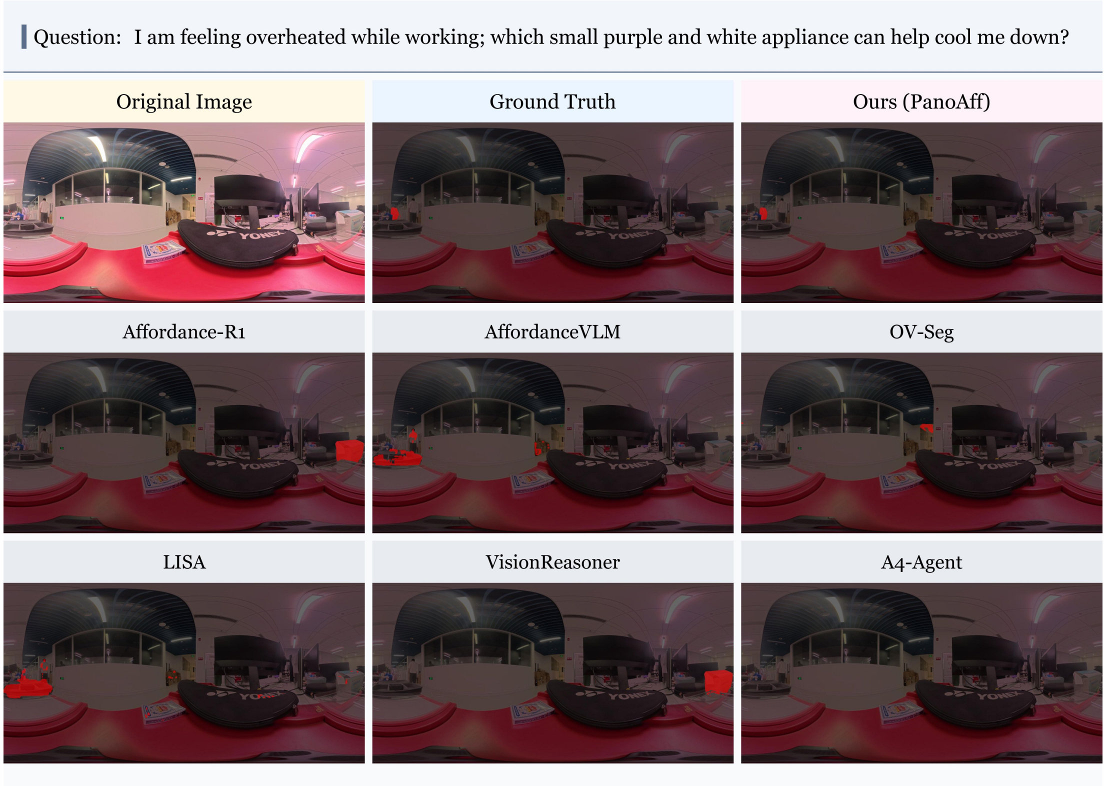
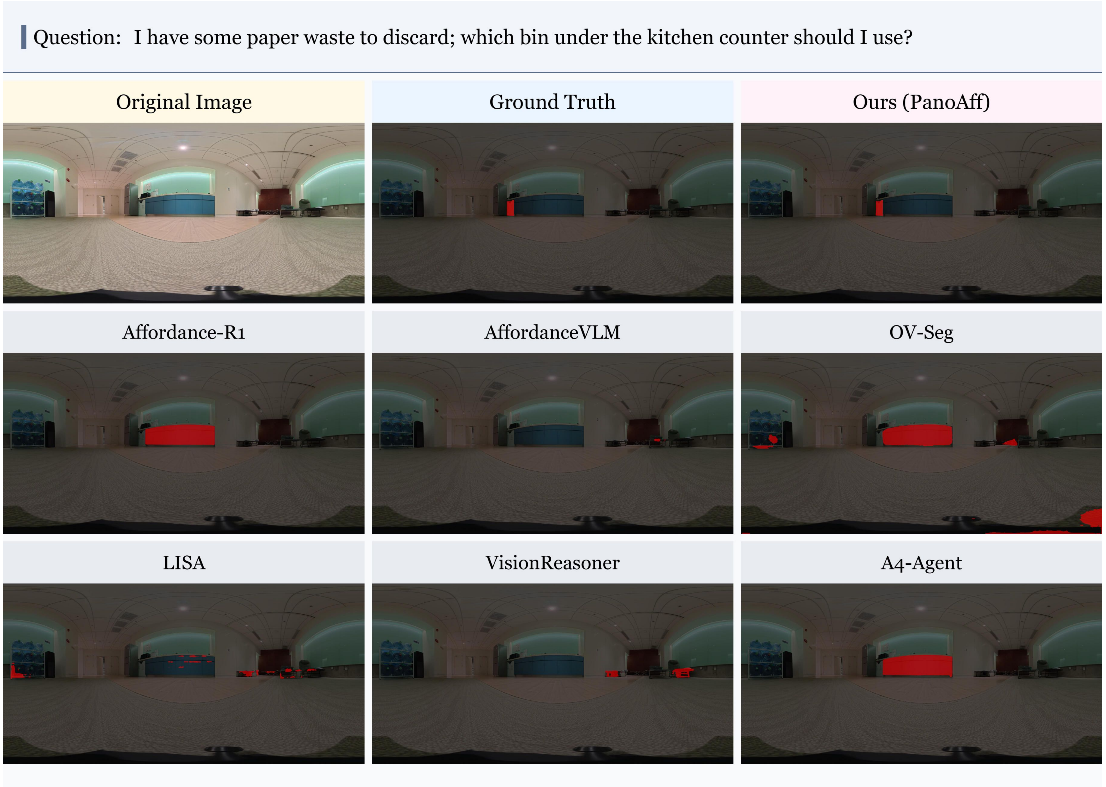

Question: My hair is wet after a shower; what device on the table can I use to dry it?

Figure 33 Qualitative comparison on PAP-12K (Livingroom).



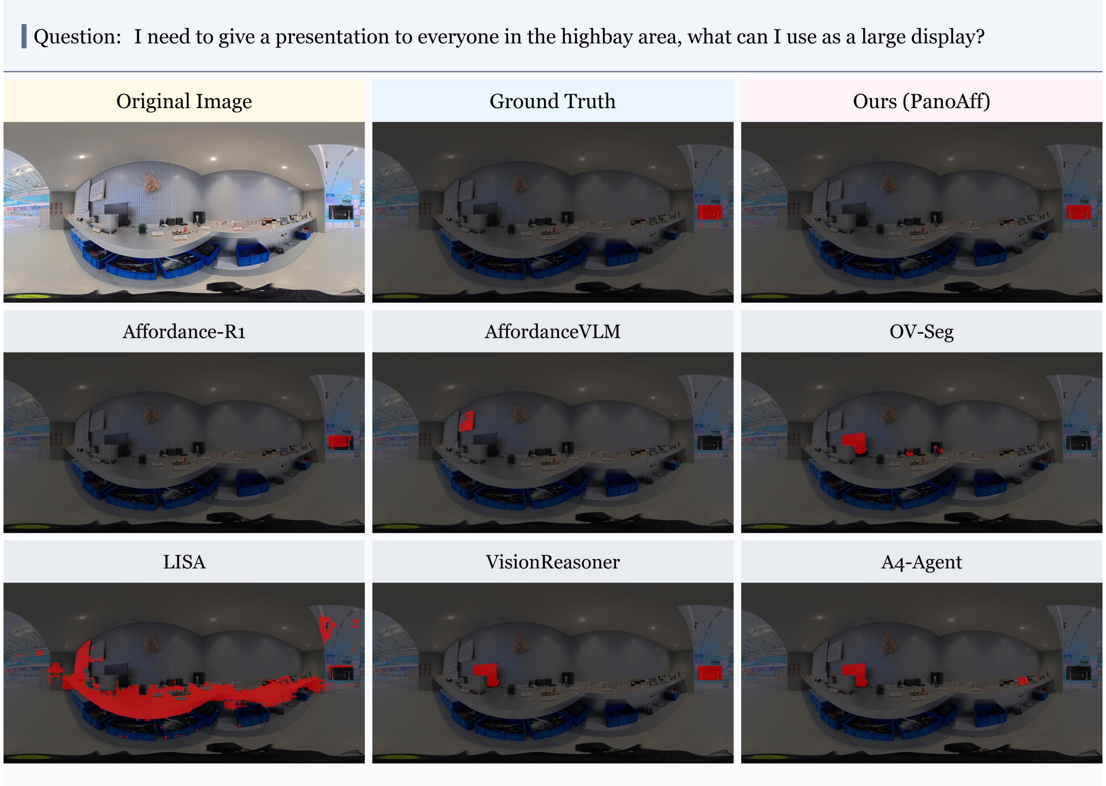
Question: I am feeling overheated while working; which small purple and white appliance can help cool me down?

Figure 34 Qualitative comparison on PAP-12K (Office).



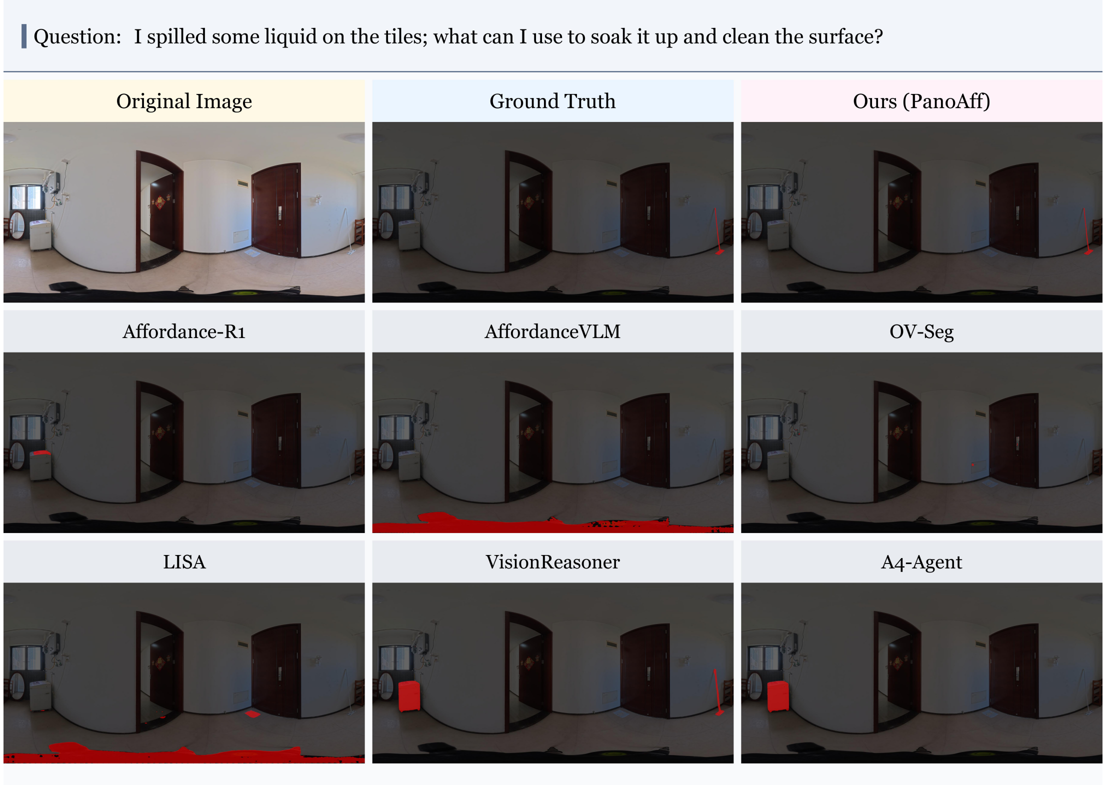
Question: I have some paper waste to discard; which bin under the kitchen counter should I use?

Figure 35 Qualitative comparison on PAP-12K (Pantry).



Question: I need to give a presentation to everyone in the highbay area, what can I use as a large display?

Figure 36 Qualitative comparison on PAP-12K (Workshop).



Question: I spilled some liquid on the tiles; what can I use to soak it up and clean the surface?

Figure 37 Qualitative comparison on PAP-12K (Others).