

Panoramic Affordance Prediction

Zixin Zhang^{1,6,†}, Chenfei Liao^{1,†}, Hongfei Zhang¹, Harold H. Chen^{1,6},
Kanghao Chen^{1,6}, Zichen Wen³, Litao Guo¹, Bin Ren⁴, Xu Zheng¹,
Yinchuan Li⁶, Xuming Hu^{1,2}, Nicu Sebe⁵, Ying-Cong Chen^{1,2,‡}

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

[†]Equal Contribution [‡]Corresponding Author

{zzhang300, cliao127}@connect.hkust-gz.edu.cn

yingcongchen@ust.hk

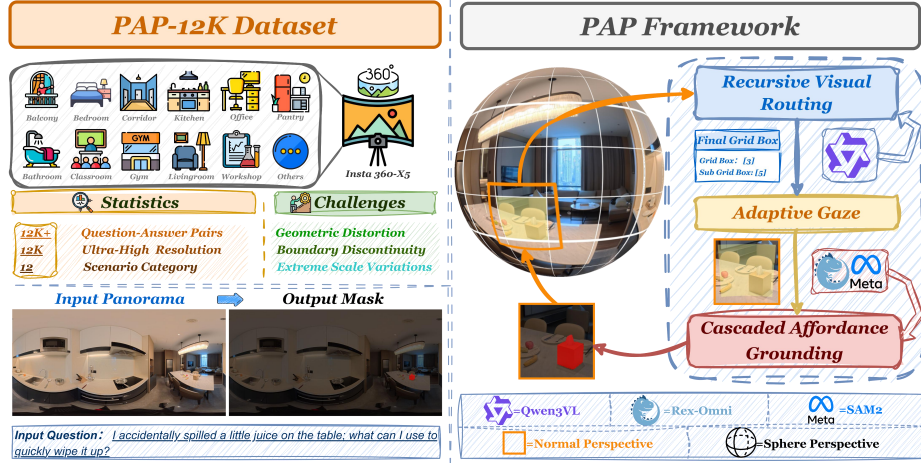


Fig. 1: Overview of our work. **Left:** We introduce PAP-12K, the first large-scale benchmark dedicated to panoramic affordance prediction, featuring ultra-high resolution (12K) imagery, rich reasoning-based QA pairs, and explicitly capturing unique panoramic challenges (geometric distortion, boundary 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.

Abstract. 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.

Keywords: Affordance · Panoramic Vision · Embodied AI

1 Introduction

Affordance [9] 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 [36, 38, 49] 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 [4, 13] and scene understanding [8, 50]. However, its potential 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. PAP mimics 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 [9] 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 [27, 39], object interaction [11, 44], and tool usage [14, 28, 48]. Consequently, it has been widely integrated into numerous robotic systems [29, 34, 41, 43]. Early learning approaches [16, 20, 26, 32] 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) [5, 15, 42], recent methods [31, 36, 38] have leveraged fine-tuning strategies to interpret complex task instructions and generate corresponding affordance maps. Most recently, A4-Agent [49] 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 [23], the advantage of “seeing everything at once” has made panoramic vision a crucial enabler for scene understanding [23, 51], virtual reality [24], autonomous driving [30], and embodied AI [35]. 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 [4, 37]. Beyond navigation, 360° visual inputs significantly enhance an agent’s ability in manipulation tasks [18] and facilitate precise localization [12]. Furthermore, integrating panoramic vision with multi-modal large language models has opened new avenues for spatial reasoning and question answering [7, 47], 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
<i>Affordance Datasets</i>						
UMD [28]	Action Category	30k	-	640 × 480	~ 50°	Real-world Capture
ReasonAff [36]	Instruction & Reasoning	3k	3k	2000 × 1500	~ 30°	Web Crawled
HANDAL [10]	Action Category	200k	-	1920 × 1440	~ 50°	Real-world Capture
3DOI [32]	Action Category	10k	-	1920 × 1080	~ 70°	Real-world Capture
RAGNet [38]	Instruction & Reasoning	273k	273k	1920 × 1440	~ 50°	Web Crawled
<i>Panoramic Datasets</i>						
SUN360 [40]	-	67.6k	-	9104 × 4552	360°	Multi-view Stitching
Matterport3D [3]	-	10.8k	-	2048 × 1024	360°	Multi-view Stitching
Stanford2D3D [1]	-	1.4k	-	4096 × 2048	360°	Multi-view Stitching
ReplicaPano [6]	-	2.7k	-	1024 × 512	360°	Synthetic Rendering
DeepPanoContext [46]	-	1.5k	-	1024 × 512	360°	Synthetic Rendering
Thinking-in-360 [45]	-	3k	-	7680 × 3840	360°	Web Crawled
Realsee3D [21]	-	10k	-	1600 × 800	360°	Multi-view Stitching
<i>Panoramic + Affordance Datasets</i>						
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 [36] and RAGNet [38]), our QA pairs are specifically designed to require complex reasoning rather than simple perception. Each pair is associated with a precise pixel-level segmentation

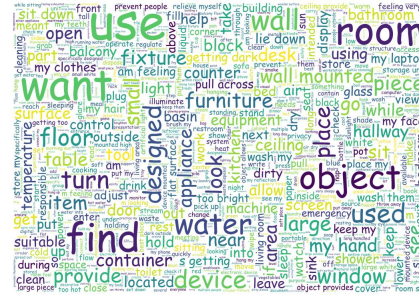


Fig. 4: Word Cloud of Questions in PAP-12K.

mask, grounding the affordance in the visual domain. This moves beyond simple classification to require explicit logical deduction and precise localization.



Fig. 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).

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 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

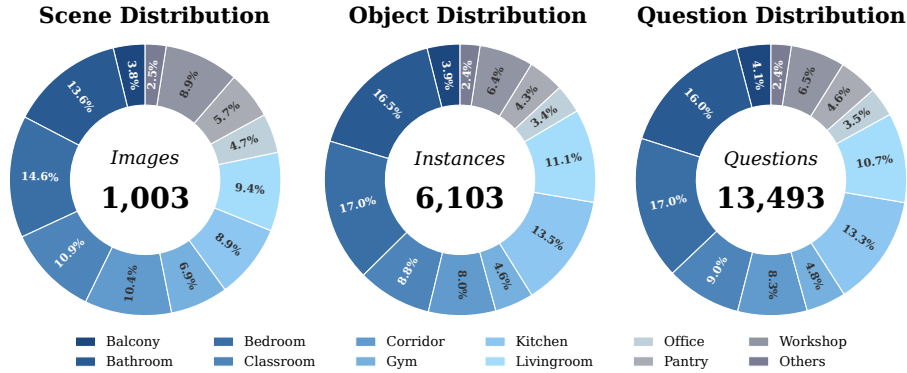


Fig.3: Statistics of the PAP-12K. (Left) Scene distribution with 1,003 high-resolution panoramic images; (Middle) Object distribution featuring 6,103 annotated object instances; (Right) Question distribution comprising 13,493 affordance questions.

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.

It is worth emphasizing that PAP-12K is **not restricted to human-centric home scenes**. Beyond domestic spaces, the 12 scene types also encompass broader embodied environments, including industrial-style settings such as **Workshop** (e.g., Fig. ??) and activity-oriented spaces such as **Gym** (which also covers outdoor-activity areas, e.g., Fig. ??). Our deliberate emphasis on indoor living spaces stems from our primary focus on embodied-agent deployment, especially home and service robots, whose interaction scenarios are predominantly indoor. We nonetheless regard the extension of PAP-12K towards more complex, cluttered industrial and outdoor environments as a valuable direction for future expansion.

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

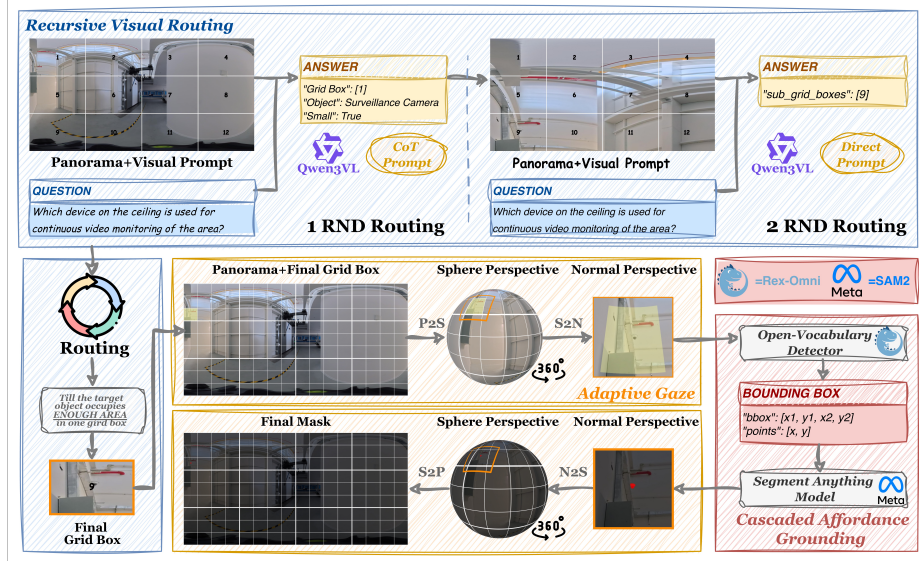


Fig. 5: Illustration of the PAP framework.

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.

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 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: **① Scale-Salient Targets**: If the target spans multiple grids, the current resolution is sufficient, and the recursion terminates. **② 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.

Table 2: Comparison on PAP-12K. **Best** results are highlighted in bold. Second-best results are highlighted in underline.

Method	gIoU $\uparrow$	cIoU $\uparrow$	P ₅₀ $\uparrow$	P ₅₀₋₉₅ $\uparrow$	Inference Time
OV-Seg [22]	29.48	17.85	32.00	18.80	$\sim 8s$
LISA [19]	15.21	16.34	13.66	8.30	$\sim 7s$
VisionReasoner [25]	49.33	44.64	51.06	38.06	$\sim 12s$
AffordanceVLM [38]	9.66	13.11	8.96	5.41	$\sim 7.8s$
Affordance-R1 [36]	51.80	<u>50.32</u>	55.47	40.70	$\sim 10.4s$
A4-Agent [49]	<u>62.55</u>	49.97	<u>67.09</u>	<u>54.28</u>	$\sim 11.8s$
PAP (Ours)	71.56	62.30	75.49	64.97	$\sim 10s$

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 $\mathcal{I}_{persp}$ 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 [49], 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 [49], Affordance-R1 [36], and AffordanceVLM [38]. Furthermore, following the evaluation protocols in [36, 49], we extend our comparison to include general Open Vocabulary Segmentation models, such as Vision Reasoner [25], OVSeg [22], and LISA [19].

Evaluation Metrics: Following the standard evaluation protocol in affordance prediction [36, 49], 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 [2] as the backbone of our VLM, and following [49], we adopt Rex-Omni [17] as the OVD model and the SAM-2-Large [33] as the Segmentation model. During the Recursive Visual Routing stage, we apply the dynamic resolution adaptation



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

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

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

Method	PAP-12K-Hard				PAP-12K-Normal			
	gIoU $\uparrow$	cIoU $\uparrow$	$P_{50}\uparrow$	$P_{50-95}\uparrow$	gIoU $\uparrow$	cIoU $\uparrow$	$P_{50}\uparrow$	$P_{50-95}\uparrow$
OV-Seg [22]	10.66	9.18	10.32	4.90	37.66	21.72	41.44	24.84
LISA [19]	4.75	13.04	3.23	2.01	19.75	17.24	18.20	11.03
VisionReasoner [25]	25.96	29.74	24.91	15.00	59.50	49.59	62.44	48.10
AffordanceVLM [38]	4.94	15.94	3.59	1.92	11.72	12.36	11.30	6.93
Affordance-R1 [36]	33.31	<u>40.84</u>	36.26	21.39	59.85	52.97	63.82	49.11
A4-Agent [49]	<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

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.

Table 4: Ablation on Prompt Style.

CoT	VGP	gIoU $\uparrow$	cIoU $\uparrow$	P ₅₀ $\uparrow$	P ₅₀₋₉₅ $\uparrow$
$\times$	$\times$	57.70	57.68	60.76	50.52
$\times$	$\checkmark$	69.56	61.05	74.02	61.90
$\checkmark$	$\times$	67.22	58.75	70.61	60.51
$\checkmark$	$\checkmark$	71.56	62.30	75.49	64.97

Table 5: Ablation on Adaptive Gaze.

Method	gIoU $\uparrow$	cIoU $\uparrow$	P ₅₀ $\uparrow$	P ₅₀₋₉₅ $\uparrow$
w/o AG	64.99	55.43	68.43	56.37
w/ AG	71.56	62.30	75.49	64.97
Δ	6.57	6.87	7.06	8.60

Table 6: Ablation on RVR.

Subset	Method	gIoU $\uparrow$	cIoU $\uparrow$	P ₅₀ $\uparrow$	P ₅₀₋₉₅ $\uparrow$
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 geometric rectification. We ablate this module by directly applying the OVD and SAM on the cropped ERP regions without Adaptive Gaze. 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 Adaptive Gaze 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.

Acknowledgements

This work was supported by the EU Horizon projects ELIAS (No. 101120237) and ELLIOT (No. 101214398) and by the FIS project GUIDANCE (No. FIS2023-03251).

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 7th IEEE International Conference on 3D Vision, 3DV 2017. pp. 667–676. Institute of Electrical and Electronics Engineers Inc. (2018)
4. Chen, Y., Pan, Y., Liu, R., Zhang, H., Zhang, G., Sun, B., Zhang, J.: 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). pp. 717–722. IEEE (2024)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
6. Dong, Y., Fang, C., Bo, L., Dong, Z., Tan, P.: Panocontext-former: Panoramic total scene understanding with a transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28087–28097 (2024)

7. Dongfang, Z., Zheng, X., Weng, Z., Lyu, Y., Paudel, D.P., Van Gool, L., Yang, K., Hu, X.: Are multimodal large language models ready for omnidirectional spatial reasoning? arXiv preprint arXiv:2505.11907 (2025)
8. Gao, S., Yang, K., Shi, H., Wang, K., Bai, J.: Review on panoramic imaging and its applications in scene understanding. *IEEE Transactions on Instrumentation and Measurement* **71**, 1–34 (2022)
9. Gibson, J.J.: The theory of affordances. *Hilldale, USA* **1**(2), 67–82 (1977)
10. Guo, A., Wen, B., Yuan, J., Tremblay, J., Tyree, S., Smith, J., Birchfield, S.: 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). pp. 11428–11435. IEEE (2023)
11. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for human-object interaction detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 495–504 (2021)
12. Huang, H., Liu, C., Zhu, Y., Cheng, H., Braud, T., Yeung, S.K.: 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. pp. 22314–22324 (2024)
13. Huang, H., Yeung, S.K.: 360vo: Visual odometry using a single 360 camera. In: 2022 international conference on robotics and automation (icra). pp. 5594–5600. IEEE (2022)
14. Huang, S., Ponomarenko, I., Jiang, Z., Li, X., Hu, X., Gao, P., Li, H., Dong, H.: 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). pp. 7580–7587. IEEE (2024)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
16. Jang, J.H., Seo, H., Chun, S.Y.: Intra: Interaction relationship-aware weakly supervised affordance grounding. In: European Conference on Computer Vision. pp. 18–34. Springer (2024)
17. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction (2025), <https://arxiv.org/abs/2510.12798>
18. Kerr, J., Hari, K., Weber, E., Kim, C.M., Yi, B., Bonnen, T., Goldberg, K., Kanazawa, A.: Eye, robot: Learning to look to act with a bc-rl perception-action loop. arXiv preprint arXiv:2506.10968 (2025)
19. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9579–9589 (2024)
20. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10922–10931 (2023)
21. Li, L., Wu, Y., Li, X., Wang, L., Rao, T., Zhou, J., Pan, C., Hui, X.: Realsee3d: A large-scale multi-view rgb-d dataset of indoor scenes (version 1.0) (2025). <https://doi.org/10.5281/zenodo.17826243>, <https://doi.org/10.5281/zenodo.17826243>
22. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In:

- Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023)
23. Lin, X., Ge, X., Zhang, D., Wan, Z., Wang, X., Li, X., Jiang, W., Du, B., Tao, D., Yang, M.H., Qi, L.: One flight over the gap: A survey from perspective to panoramic vision. *arXiv preprint* (2025)
 24. Lin, X., Luo, S., Shan, X., Zhou, X., Ren, C., Qi, L., Yang, M.H., Vasconcelos, N.: Hqgs: High-quality novel view synthesis with gaussian splatting in degraded scenes. In: *ICLR* (2025)
 25. Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081* (2025)
 26. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning affordance grounding from exocentric images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 2252–2261 (2022)
 27. Mo, K., Guibas, L.J., Mukadam, M., Gupta, A., Tulsiani, S.: Where2act: From pixels to actions for articulated 3d objects. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6813–6823 (2021)
 28. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: *2015 IEEE international conference on robotics and automation (ICRA)*. pp. 1374–1381. IEEE (2015)
 29. Nasiriany, S., Kirmani, S., Ding, T., Smith, L., Zhu, Y., Driess, D., Sadigh, D., Xiao, T.: Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 8249–8257. IEEE (2025)
 30. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with kins dataset. In: *CVPR* (2019)
 31. Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., Li, L.E.: Affordancellm: Grounding affordance from vision language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7587–7597 (2024)
 32. Qian, S., Fouhey, D.F.: Understanding 3d object interaction from a single image. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 21753–21763 (2023)
 33. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024)
 34. Singh, A., Torshizi, K., Habib, K., Yu, K., Gao, R., Tokekar, P.: Afford2act: Affordance-guided automatic keypoint selection for generalizable and lightweight robotic manipulation. *arXiv preprint arXiv:2510.01433* (2025)
 35. Wan, Z., Bi, Z., Zhou, Z., Ren, H., Zeng, Y., Li, Y., Qi, L., Yang, X., Yang, M.H., Cheng, H.: Rapid hand: A robust, affordable, perception-integrated, dexterous manipulation platform for generalist robot autonomy. *arXiv preprint arXiv:2506.07490* (2025)
 36. Wang, H., Wang, S., Zhong, Y., Yang, Z., Wang, J., Cui, Z., Yuan, J., Han, Y., Liu, M., Ma, Y.: Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. *arXiv preprint arXiv:2508.06206* (2025)
 37. Wang, S., Zhou, D., Xie, L., Xu, C., Yan, Y., Yin, E.: Panogen++: Domain-adapted text-guided panoramic environment generation for vision-and-language navigation. *Neural Networks* **187**, 107320 (2025)

38. Wu, D., Fu, Y., Huang, S., Liu, Y., Jia, F., Liu, N., Dai, F., Wang, T., Anwer, R.M., Khan, F.S., et al.: Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11980–11990 (2025)
39. Wu, R., Zhao, Y.: Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. In: International Conference on Learning Representations (ICLR), 2022 (2022)
40. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2695–2702. IEEE (2012)
41. Xu, R., Zhang, J., Guo, M., Wen, Y., Yang, H., Lin, M., Huang, J., Li, Z., Zhang, K., Wang, L., et al.: A0: An affordance-aware hierarchical model for general robotic manipulation. arXiv preprint arXiv:2504.12636 (2025)
42. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
43. Ye, A., Zhang, Z., Wang, B., Wang, X., Zhang, D., Zhu, Z.: Vla-r1: Enhancing reasoning in vision-language-action models. arXiv preprint arXiv:2510.01623 (2025)
44. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22479–22489 (2023)
45. Yu, H., Han, Y., Zhang, X., Yin, B., Chang, B., Han, X., Liu, X., Zhang, J., Pavone, M., Feng, C., et al.: Thinking in 360 $\{\deg\}$: Humanoid visual search in the wild. arXiv preprint arXiv:2511.20351 (2025)
46. Zhang, C., Cui, Z., Chen, C., Liu, S., Zeng, B., Bao, H., Zhang, Y.: 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. pp. 12632–12641 (2021)
47. Zhang, X., Ye, Z., Zheng, X.: Towards omnidirectional reasoning with 360-r1: A dataset, benchmark, and grpo-based method. arXiv preprint arXiv:2505.14197 (2025)
48. Zhang, Z., Chen, K., Lin, X., Jiang, L., Zheng, X., Lyu, Y., Guo, L., Li, Y., Chen, Y.C.: Phystoolbench: Benchmarking physical tool understanding for mllms. arXiv preprint arXiv:2510.09507 (2025)
49. Zhang, Z., Chen, K., Wang, H., Zhang, H., Chen, H.H., Liao, C., Guo, L., Chen, Y.C.: A4-agent: An agentic framework for zero-shot affordance reasoning. arXiv preprint arXiv:2512.14442 (2025)
50. Zheng, X., Liao, C., Weng, Z., Lei, K., Dongfang, Z., He, H., Lyu, Y., Jiang, L., Qi, L., Chen, L., Paudel, D.P., Yang, K., Zhang, L., Van Gool, L., Hu, X.: Panorama: The rise of omnidirectional vision in the embodied ai era. arXiv preprint (2026)
51. Zhou, Y., Zhang, T., Zhang, D., Ji, S., Li, X., Qi, L.: Dense360: Dense understanding from omnidirectional panoramas. arXiv preprint arXiv:2506.14471 (2025)

Panoramic Affordance Prediction

Supplementary Material

Contents