QueryMe: Query-Driven Open-Vocabulary 3D Object Affordances Grounding from Multimodal Evidence

Weiyu Zhao[1], Ru Li[1], Jiaqi Liu[1], Sizhe Zhao[1], Qinglin Liu[1], Shengping Zhang[†,1,2]
[1]Harbin Institute of Technology,
[2]Harbin Institute of Technology (Weihai) Qingdao Research Institute
weiyuzhao66@gmail.com, liru@hit.edu.cn
{2024213059, sizhe.zhao}@stu.hit.edu.cn,{qlliu, s.zhang}@hit.edu.cn

Abstract

*Open-vocabulary 3D object affordance grounding aims to identify functional regions of objects given arbitrary semantic descriptions. However, existing methods often rely on fixed training categories and geometric priors, lacking geometric invariance and analogical reasoning capabilities. Since there exists a significant domain gap when transferring affordance knowledge learned from 2D images to 3D point clouds, existing methods struggle to generalize well to objects with diverse shapes or unseen categories, and fail to perform effective category reasoning. To address these challenges, we propose **QueryMe**, a **Query**-driven framework that learns from **M**ultimodal **e**vidence spaces to achieve open-vocabulary 3D affordance grounding. The proposed approach is to project human-object interaction images into 3D space, employ an Adaptive Spatial Attention module to focus on key interaction regions, and introduce a multimodal query structure to retrieve available geometrically consistent functional parts within the point cloud, effectively fusing visual, linguistic, and geometric cues. Leveraging attention-based query mechanisms, our method adaptively localizes affordance regions and performs analogy reasoning through geometric similarity, thereby exhibiting strong generalization to unseen scenes and objects. Experimental results demonstrate that QueryMe consistently outperforms state-of-the-art approaches, with the AUC improving by 4.19% compared to previous work for unseen affordance grounding tasks.*

1. Introduction

Humans inherently acquire the ability to use tools by observing others and recognizing the functional regions of objects. For instance, one can intuitively avoid the sharp edge of a blade while exploiting it for cutting. Inspired by this capability, Open-Vocabulary 3D Object Affordance Grounding [40] seeks to equip robots with the ability to localize and interpret functional regions of objects based on their affordances. This task enables embodied intelligence to understand and interact with the physical world, and finds broad applications in robotic manipulation [9, 18], Vision-Language-Action (VLA) models [14, 22], and imitation learning [15].

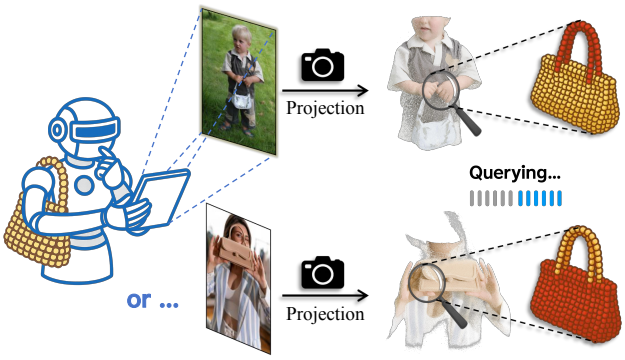

Figure 1. **QueryMe** localizes object affordances from multimodal evidence. From a single HOI image, it projects to an 3D space and performs query-driven cross-modal reasoning over HOI, text, and object features to mark affordance regions on the target object, aligning diverse interaction priors, narrowing the 2D to 3D gap, and generalizing to unseen scenes and categories.

Recent studies [17, 41] have attempted to combine the geometric structure of point clouds with semantic labels to enable affordance detection in 3D environments. However, these methods typically rely on object categories and geometric structures present in the training data, resulting in poor generalization when encountering previously unseen objects. This limitation restricts their scalability and practical deployment in open-world settings. To mitigate this issue, subsequent research has incorporated additional information sources, such as textual descriptions [17, 39], Human-Object Interaction (HOI) images [41], 2D projections of point clouds [19], and more detailed 3D geometric

† Corresponding author (s.zhang@hit.edu.cn)

information [13, 38]. By exploring these modalities, models gain stronger priors for reasoning about functional regions, thereby improving their ability to generalize to unseen objects and affordance locations. While these methods have made progress to some extent, they still face several challenges, such as the lack of effective geometric invariance modeling, especially for shape differences among similar objects, and the absence of analogical reasoning capabilities for the diverse usage modes of the same object.

Most recent works have started leveraging prior knowledge from Vision-Language Models (VLM) to alleviate the geometric discrepancies introduced by the diversity of object affordances. Techniques such as utilizing pre-trained models [44], incorporating chain-of-thought [29], and applying model distillation [2] have partially mitigated these issues. While these approaches demonstrate the potential of VLM priors in bridging affordance diversity, they remain limited in several ways. First, existing methods often fall short of fully integrating multimodal information, typically relying on either visual or textual cues in isolation rather than a coherent cross-modal representation. Second, when affordance learning is conducted directly from 2D HOI images, substantial domain gaps arise between 2D image distributions and 3D object affordances. These domain gaps hinder reliable translation to 3D environments and restrict model performance in more complex scenarios. Thus, effectively combining multimodal priors remains an open challenge for robust and scalable 3D affordance grounding.

Findings in cognitive psychology suggest [28, 43] that human object recognition progresses from general geometric shapes to higher-level functional attributes. Following this progression, we introduce a 3D query mechanism that, conditioned on the object, retrieves the geometric areas that humans typically interact, rather than explicitly pairing objects with predefined regions. This process is abstracted as affordance grounding. As shown in Fig. 1, explicitly learning object geometry and functional behavior from reconstructed 3D point clouds can effectively alleviate the challenges of cross-domain learning. In unseen scenes, exploiting geometric similarity and attending to hand–object interaction regions enhances the generalization of affordance grounding. Meanwhile, with the rapid progress of monocular reconstruction [34, 36], mapping 2D HOI images into a 3D feature space has become feasible. Building on the above findings, we introduce **QueryMe**, a novel framework that effectively queries and grounds 3D object affordances from multimodal evidence, including images of human–object interactions and natural-language instructions.

Specifically, we first map an HOI image into a 3D scene using a feed-forward 3D reconstruction pipeline. To mitigate sensitivity to reconstruction inaccuracies and background clutter, we introduce a *Adaptive Spatial Attention Module* that selectively suppresses irrelevant pixels. On the reconstructed 3D object point cloud, we apply random spatial sampling to construct a compact set of learnable affordance query vectors, which substantially reduces computation and stabilizes optimization. Building upon this representation, we propose a *Multimodal Guided Query Learning Module* that sequentially retrieves affordance cues across three domains: the 3D HOI space, the text space, and the 3D object space. Finally, we employ a lightweight query decoder that takes the fused 3D object HOI representation as input and localizes the affordance region of operation.

To summarize, our contributions are the following:

- We propose **QueryMe**, an effective framework that queries information from multi-modal inputs and advances 3D object affordance grounding.
- We demonstrate that mapping human–object interaction regions into 3D space enables effective extraction of affordance knowledge, alleviating the domain discrepancies between 2D images and 3D object point clouds and improving generalization to unseen scenes.
- We design an attention-based query mechanism that adaptively localizes affordance operation regions, making the model robust across diverse task settings.
- We conduct extensive experiments showing clear gains over strong baselines, with particularly pronounced improvements on unseen-scene splits, underscoring the practicality of our approach.

2. Related Works

2.1. 2D Affordance Grounding

2D affordance grounding localizes actionable regions in images, identifying where interactions such as grasping or pushing can occur. Early studies established the concept using handcrafted cues, geometric reasoning, and shape priors [8]. With deep learning, CNN-based approaches became dominant, enabling pixel- and object-level affordance segmentation [23, 27]; AffordanceNet exemplifies an effective end-to-end design [5]. Nevertheless, both classical pipelines and CNN-based dense segmentation remain tied to closed affordance sets, limiting generalization to unseen objects [5, 12]. Recent language-conditioned models address this issue by enabling open-vocabulary grounding with stronger semantic generalization [2, 22].

To relax the closed-set constraint, language-guided and task-conditioned approaches emerged, bridging visual perception with semantic understanding. These methods leverage linguistic priors to extend grounding beyond rigid label sets, with applications ranging from human intention understanding via gaze [35] to social interaction modeling [30]. Most recently, foundation models have enabled open-vocabulary affordance prediction. Key enablers include CLIP [26] for large-scale vision-language embedding and SAM [11] for promptable universal segmentation.

Building on these, affordance-specific systems such as AffordanceLLM [25] and hybrid designs like BiT-Align [10] integrate vision-language model (VLM) reasoning with geometric cues to achieve more robust grounding. Despite these advances, integrating task intent, semantic knowledge, and fine-grained spatial structures remains non-trivial, leaving open issues in compositional generalization and environmental robustness.

2.2. 3D Affordance Grounding

Early research [20, 32] used geometric features and manual annotations to map function from shape. Such geometry-based methods worked on known objects but generalized poorly when similar shapes implied different uses.

To overcome these limitations, researchers pivoted toward data-driven approaches fueled by large-scale annotated datasets. This shift significantly advanced 3D affordance perception. For example, AffordanceNet-3D [4] provides extensive object and part level affordance annotations, enabling supervised learning at scale. Complementary efforts have both learned structured affordance spaces for robotic manipulation [39], highlighting the benefits of compositional representations, and applied deep reinforcement learning for affordance reasoning in robotic control [21]. To move beyond closed sets of affordance labels, early attempts introduced end-to-end training for manipulation-oriented affordance learning [6], followed by differentiable pipelines that improved optimization and scalability [16]. Subsequently, several studies leveraged human–object interaction images [41] and textual priors [17] to guide open-vocabulary affordance grounding, further enhancing generalization to unseen objects and actions. Geometry-aware generative frameworks render photorealistic 3D scenes and enable fine-grained affordance segmentation, exemplified by SeqAffordSplat [13] and 3DAffordSplat [37]. DAG [33] employs a diffusion model to map representations learned from HOI images to corresponding 3D affordance regions, enabling fine-grained grounding of interaction-relevant areas. 3D-AffordanceLLM [2] and GREAT [29] marry LLMs with geometric perception, achieving open-vocabulary affordance grounding in complex, realistic scenes.

Despite notable progress, existing approaches remain constrained by fixed training taxonomies and hand-crafted geometric priors, exhibiting limited geometric invariance and weak analogical reasoning. A substantial domain gap in transferring HOI knowledge from 2D images to 3D point clouds impedes generalization to diverse shapes and unseen categories while hampering category-level reasoning. To address this, we propose **QueryMe**, which learns geometry priors from 3D HOI space and introduces a multimodal evidence–query mechanism to improve generalization.

3. Methods

3.1. Overview

We input the object point cloud $P \in \mathbb{R}^{N\times 3}$ and a single Human-Object Interaction (HOI) RGB image $I \in \mathbb{R}^{3\times H\times W}$, where N is the number of points, and H and W represent the height and width of the image. Following the setup in GREAT[29], we obtain the text representations of Interaction Attributes and Geometric Attributes, denoted as $T = \{T_i, T_g\}$, via the Chain-of-Thought from the Vision-Language Model (VLM). Additionally, we employ a feed-forward 3D reconstruction method[34] to map the image I into the 3D space, resulting in $H \in \mathbb{R}^{N'\times 6}$. The goal is to optimize the model M to output the object affordance φ, where $\varphi = M(H, T, P)$. To make this prediction both efficient and robust, Sec. 3.2 introduces an Auto-Adaptive Spatial Anchor Selection module that identifies salient anchors in the reconstructed HOI space and concentrates subsequent computation on interaction-relevant regions; building on these anchors, Sec. 3.3 performs Multimodal Feature Encoding to align and fuse textual attributes, 3D HOI features, and object point-cloud representations into a unified embedding; Sec. 3.4 then applies Multimodal Guided Query Learning, which leverages the fused embedding to steer learnable queries toward affordance-relevant locations and refine them via attention; finally, Sec. 3.5 details the Affordance Decoder and the training objectives that transform query–point interactions into a point-wise affordance map and optimize the model with focal and Dice losses.

3.2. Adaptive Spatial Attention Module

In reconstructing HOI images into 3D space, feed-forward 3D reconstruction methods often introduce inherent noise and retain large amounts of irrelevant background information. Consequently, extracting compact and clean geometric features in the reconstructed 3D HOI space becomes essential. When processing high-resolution point cloud data, directly handling the full set of points leads to significant computational overhead and excessive memory consumption. To address this, while preserving the integrity of spatial information and improving computational efficiency, we propose a lightweight Auto-Adaptive Spatial Anchor strategy for importance-aware spatial feature extraction.

We first sample a subset of points $\mathcal{P}_s = \{p_j\}_{j=1}^{N_s}$ from the input point cloud $\mathcal{P}$ using a global sampling strategy, retaining a fraction p of the points (so $N_s = pN$). A lightweight MLP $\Phi : \mathbb{R}^3 \to \mathbb{R}^d$ then encodes each sampled point's coordinates into a d-dimensional feature vector. Next, to capture spatial dependencies among these sampled points without explicitly constructing a graph, we feed the sequence of feature vectors into a 1D convolutional layer. This allows the network to model spatial continuity and local topology along the sequence of sampled points.

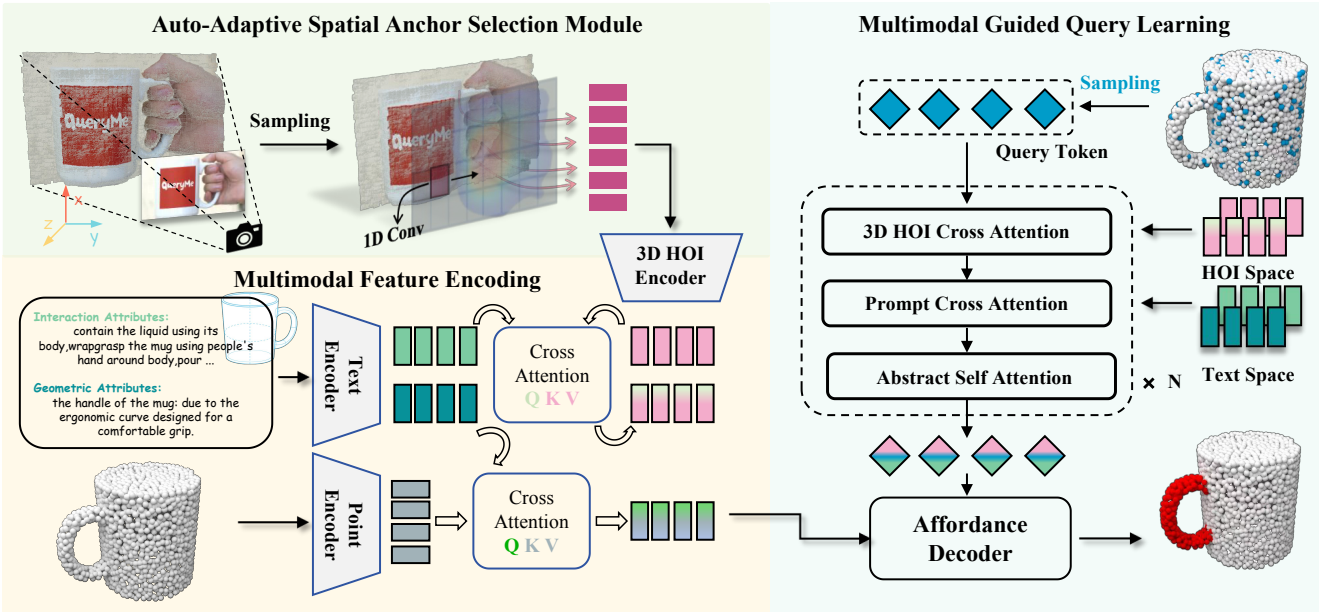

Figure 2. **Overview of the QueryMe framework.** (A) ***Adaptive Spatial Attention Module***, which reconstructs HOI images into 3D and applies a 1D convolution to extract 3D interaction features, thereby mapping 2D interaction cues into the 3D space; (B) ***Multimodal Feature Encoding***, which employs a point-cloud encoder and a text encoder to extract geometric and linguistic features and aligns them via cross-modal attention. (C) ***Multimodal Guided Query Learning***, which introduces learnable query tokens and fuses evidence from the 3D HOI, text, and object spaces through cross-modal attention, followed by an affordance decoder to predict the interaction region. The framework highlights 2D-to-3D cross-domain mapping, cross-modal feature fusion, and spatial alignment.

After the convolutional feature aggregation, an importance predictor assigns each sampled point p_j an importance score s_j. These scores indicate the geometric and semantic salience of the sampled points, guiding the network to focus on important structures. Finally, we propagate the importance scores from the sampled subset back to the entire point cloud via distance-based interpolation. For each original point $p_i \in \mathcal{P}$, we first compute the distance-based weight w_{ij} between it and each sampled point, and then interpolate the importance scores based on these weights. Specifically, the weight w_{ij} is computed as:

$$w_{ij} = \frac{1}{|p_i - p_j|^2 + \epsilon} \tag{1}$$

where ϵ is a small constant to ensure numerical stability. Then, the interpolated importance score $\hat{s}_i$ for the original point p_i is computed as a weighted average of the sampled points' importance scores:

$$\hat{s}_i = \frac{\sum_{j=1}^{N_s} w_{ij} s_j}{\sum_{j=1}^{N_s} w_{ij}} \tag{2}$$

where w_{ij} is the weight based on the distance between the original point q_i and the sampled point p_j, and $\hat{s}_i$ is the interpolated importance score for q_i. This interpolation scheme encourages nearby points to have similar importance values, enforcing local smoothness and preserving structural consistency across the point cloud. By adaptively selecting a representative set of anchor points and spreading their influence to neighboring points, our module highlights geometrically and semantically salient structures, which can improve the efficiency and effectiveness of downstream tasks.

3.3. Multimodal Feature Encoding

Text Feature Encoder. We employ two RoBERT models with identical architectures to separately encode the interaction attributes T_i and the geometric attributes T_g, using independently learned parameters. To capture their mutual dependencies in the textual feature space, we insert a lightweight bidirectional cross-attention module (i.e., T_i attends to T_g and vice versa) in each branch, followed by a shallow projection head. The refined representations are denoted as $T' = \{T_i', T_g'\}$.

3D HOI Encoder. As shown in Sec. 3.2, we design a 3D HOI encoder to extract hierarchical spatial features guided by the predicted importance scores. Given the importance score $\hat{s}_i$ for each point p_i in the reconstructed point cloud, we first select the top-k points to form a candidate set:

$$P_k = \{\, p_i \mid \mathrm{rank}(\hat{s}_i) \leq k \,\} \tag{3}$$

From this set, we uniformly sample M_p points and arrange

them into a matrix $\mathbf{P} \in \mathbb{R}^{3\times M_p}$, which encodes the 3D coordinates of the selected points. To effectively capture their geometric information, we employ a two layers PointNet++ [24], which performs hierarchical sampling and feature aggregation. The first layer models fine-grained local structures, while the second captures high level spatial context, resulting in a compact yet expressive representation:

$$\mathbf{H'}_{\mathrm{o}} = \Phi_{\mathrm{HOI}}(\mathbf{P}) \in \mathbb{R}^{D\times M_f} \tag{4}$$

This hierarchical design preserves both detailed local geometry and global interaction structure. We fuse the multiscale point cloud features $\mathbf{H'}_o$ with the textual interaction features T'_i through a single cross-attention layer. T'_i serves as the query vector, while $\mathbf{H'}_o$ provides the key and value features, yielding the interaction-aware representation:

$$\mathbf{H'}_{\mathrm{i}} = \mathrm{CrossAttn}\big(\mathbf{T'}_i,\ \mathbf{H'}_{\mathrm{o}}\big) \tag{5}$$

The overall HOI latent representation is then defined as:

$$\mathbf{H'} = \{\mathbf{H'}_{\mathrm{o}},\ \mathbf{H'}_{\mathrm{i}}\} \tag{6}$$

Point Encoder. To better process the object point cloud P, we adopt a hierarchical PointNet++ backbone that captures multi scale geometry and broader spatial context, improving robustness to noise and varying sampling density. The intermediate features P' are passed through a lightweight projection and normalization, then serve as the key and value inputs to the abstract self attention module described in Sec. 3.4. To further align semantics with geometry, we fuse the deepest encoder features with T'_g via cross attention, followed by a residual MLP refinement, which enhances feature alignment and integration and yields the final point cloud representation P'.

3.4. Multimodal Guided Query Learning

We propose a decoder that leverages query learning to model the interaction between object affordance grounding and prompt information. We first sample a set of locations from the input object point cloud P using farthest point sampling and record their spatial coordinates *pos*. A set of learnable query vectors is initialized to zero, and positional information is injected via an MLP positional encoder $\phi_{\mathrm{pos}} : \mathbb{R}^3 \to \mathbb{R}^d$ applied to *pos*. Empirically, we observed that 3D ROPE [31] provides limited gains on a fixed single point cloud, likely because rotary phase coupling is less informative when the geometry does not vary across frames, whereas the MLP encoder offers stronger locality and smoother optimization in this setting. We set the number of decoder queries to match the number of FPS-selected points, ensuring a one-to-one correspondence between them. Specifically, each query $q_0 \in \mathbb{R}^{K\times d}$ corresponds to a position $pos \in \mathbb{R}^{K\times 3}$. The decoder attends to a multimodal memory $M = \{T', H', P'\}$ and applies cross attention in the fixed order $T' \to H' \to P'$ at every layer. This schedule progressively injects information, first textual priors, then 3D human object interaction cues, and finally intermediate point cloud features, so that the model learns affordances and interaction locations from coarse to fine. Finally, we reinject positional information into the MLP enhanced queries and compute self attention to abstract the object's intrinsic geometric structure. Residual connections and layer normalization follow each attention layer, as shown in Alg. 1.

Algorithm 1: Query Learning Mechanism

Input: Query q_0, Modality features T', H', P', spatial coordinates *pos*, Number of layers n
Output: Final query representation $q^{(n)}$

Initialize $q \leftarrow q_0$;
$q \leftarrow q + \phi_{\mathrm{pos}}(pos)$;
for *each layer* $l = 1, 2, \ldots, n$ **do**
  for *each modality* $m \in \{T', H', P'\}$ **do**
    $q_m^l \leftarrow \texttt{CrossAttention}(q, m)$;
    $q_m^l \leftarrow \texttt{LayerNorm}(q + q_m^l)$;
    $q_m^l \leftarrow \texttt{FFN}(q_m^l)$;
    Update query $q \leftarrow q_m^l$;
  $q^l \leftarrow q^l + \phi_{\mathrm{pos}}(pos)$;
  $q^l \leftarrow \texttt{SelfAttention}(q^l, q^l)$;
  Update query $q \leftarrow q^l$;
return q;

3.5. Affordance Decoder and Loss

We learn the query vector q through the query learning mechanism. To facilitate the fusion of multimodal features, we introduce a query attention mechanism that computes the interaction between the query vector q and the point cloud features P'. The resulting fused features are processed as follows:

$$\varphi = \sigma(f[\mathrm{Attention}(q,\ P')]) \tag{7}$$

where f represents the function used to compute the feature representation and final output, σ denotes the sigmoid activation function, and $\varphi \in \mathbb{R}^{N\times 1}$ corresponds to the 3D affordance of the object.

Following previous works [29], we focus on the difference between the 3D object affordance φ and the ground truth affordance label P_{label}, allowing the model to directly link the 3D object affordance with interaction images through reasoning. The total loss is composed of the focal loss and Dice loss, which supervise the point-wise heatmaps, and is formulated as:

$$L_{\mathrm{total}} = L_{\mathrm{focal}} + L_{\mathrm{dice}} \tag{8}$$

This design enables the model to learn accurate 3D affordance region representations through multimodal feature fusion and attention mechanisms, without the need for explicit supervision of affordance categories.

4. Experiments

4.1. Dataset

We train our models on PIADv2 [29]. Its point-cloud corpus is a mixture of samples from 3DIR [42], 3D-AffordanceNet [4], and Objaverse [3], covering 43 object categories and 24 affordance categories. Following the evaluation protocol in GREAT [29] and LASO [17], we adopt three standard splits: **Seen**, where training and test sets share the same object and affordance categories to evaluate in-distribution performance; **Unseen Object**, where affordance categories are shared but some object categories are held out during training to assess transfer to novel objects under known affordances; and **Unseen Affordance**, where object categories may overlap but certain affordance categories are excluded from training to test zero-shot generalization to novel affordances.

4.2. Implementation Details

All experiments are conducted on two NVIDIA L20 GPUs. We train for 50 epochs with a batch size of 8 and a learning rate of 1×10^{-5}. For geometry extraction, we employ VGGT [34] to obtain 3D information from 2D images, and we use PointNet++ [24] as the 3D backbone. We compare our method with two cross-modal learning approaches, FRCNN [40]and XMF [1], as well as three leading works in 3D open-vocabulary affordance grounding, including LAG [41], LASO [17], and GREAT [29]. In addition, we construct a baseline by directly concatenating the HOI and Text features shown in Fig. 2 into the decoder of our framework. Based on the above, we aim to address the following research questions:

- **Q1:** How does QueryMe perform compared to other baseline methods?
- **Q2:** Under the proposed Multimodal Evidence Query mechanism, how does each modality contribute to the final affordance grounding performance?
- **Q3:** Can monocular feed-forward 3D reconstruction improve performance in *unseen* scenarios?

Evaluation Metrics. To comprehensively assess the proposed method, we evaluate it and several baselines using four metrics: AUC, aIoU, SIM, and MAE.

4.3. Main Result

As shown in Tab. 1, **QueryMe** achieves superior performance across most metrics under all 3 evaluation settings. Furthermore, qualitative results illustrated in Fig. 3 providing an intuitive comparison of grounding quality across different scenarios. In the following, we present a more detailed analysis and comparison with existing baselines.

QueryMe vs. Other Baselines. Compared with the runner-up method GREAT [29], our approach achieves consistent improvements across all metrics in both the Seen and Unseen Affordance settings. In the Unseen Object setting, QueryMe outperforms GREAT [29] by 3.46%, 1.60%, and 0.02% in terms of AUC, aIOU, and SIM, respectively. Although the MAE shows a slight degradation compared with GREAT [29], it remains lower than other baselines such as LASO [17]. We attribute this minor increase to our query mechanism, which assigns small activation scores to a few non-affordance points near affordance regions. Given that each point cloud contains only 2048 points, this subtle effect becomes amplified in the MAE computation, but it does not affect the overall grounding accuracy. As shown in Fig. 3, we present the visualization results for rigidly rotated objects, planar structures, and elongated structures under different settings. Our method effectively handles 3D object point clouds with varying geometric configurations. Additionally, as shown in Fig. 4, compared to previous work [29], our approach can accurately ground the geometric location of the object, avoiding over localization. Even in cases where an object has multiple affordance regions, our method is capable of simultaneous localization, demonstrating its superior performance.

Unseen Performance. A noticeable performance degradation can be observed across all previous methods under the Unseen settings. While approaches such as IAG [41], LASO [17], and GREAT [29] progressively improve generalization by better integrating visual and textual cues, a clear gap remains when encountering unseen affordance regions. In particular, under the Unseen Affordance setting, our method surpasses GREAT [29] by 4.19% in terms of AUC. This demonstrates that QueryMe effectively alleviates the generalization limitation observed in prior works. Such improvement is attributed to projecting HOI images into 3D space and employing a multimodal evidence–driven query learning mechanism. By learning the geometric structures of hand–object interactions within the 3D HOI space, our model can abstract and generalize affordance regions that have not been observed during training. Based on this, **we can answer Q1:** QueryMe outperforms baseline methods and generalizes better to unseen settings.

4.4. Ablation Studies

We conducted a comprehensive ablation study to validate the effectiveness of our model design. The ablation experiments are divided into two parts:

Multimodal Query Mechanism. To probe how query learning harvests affordance evidence from different modal spaces, we conduct a modality ablation in which we remove each input stream in turn (In Tab. 2). When all three fea-

Table 1. **Main Results.** Evaluation metrics of comparison methods on the PIADv2 [29] dataset. Seen, Unseen Object and Unseen Affordance are three partitions of the dataset. Best values are in **bold**; ↑ indicates higher is better and ↓ indicates lower is better.

Methods	Seen				Unseen Object				Unseen Affordance			
	AUC↑	aIOU↑	SIM↑	MAE↓	AUC↑	aIOU↑	SIM↑	MAE↓	AUC↑	aIOU↑	SIM↑	MAE↓
FRCNN [40]	87.05	33.55	0.600	0.082	72.20	18.08	0.362	0.152	59.08	7.96	0.210	0.156
XMF [1]	87.39	33.91	0.604	0.078	74.61	17.40	0.361	0.126	60.99	8.11	0.225	0.152
IAG [41]	89.03	34.29	0.623	0.076	73.03	16.78	0.351	0.123	62.29	8.99	0.251	0.141
LASO [17]	90.34	34.88	0.627	0.077	73.32	16.05	0.354	0.123	64.07	8.37	0.228	0.140
GREAT [29]	91.99	38.03	0.676	0.067	79.57	20.16	0.402	**0.109**	69.81	12.05	0.290	0.127
QueryMe(Ours)	**92.34**	**39.39**	**0.683**	**0.061**	**83.03**	**21.76**	**0.420**	0.118	**74.00**	**13.76**	**0.316**	**0.097**

Table 2. **Ablation on Multimodal Evidence.** We evaluate QueryMe by selectively enabling object points features P' , 3D HOI features H', and text prompts T'. A checkmark ✓indicates the modality is used, and a cross ✗ indicates it is removed. We report AUC, aIoU, SIM, and MAE on the *Seen*, *Unseen-Object*, and *Unseen-Affordance* splits of PIADv2 [29]. Using all three modalities yields the best performance, best results are shown in **bold**.

Multimodal Ev			Seen				Unseen Object				Unseen Affordance			
Obj	**HOI**	**Text**	AUC↑	aIoU↑	SIM↑	MAE↓	AUC↑	aIoU↑	SIM↑	MAE↓	AUC↑	aIoU↑	SIM↑	MAE↓
✗	✗	✗	88.52	33.08	0.623	0.084	77.49	19.09	0.399	0.128	67.42	13.07	0.291	0.137
✓	✗	✗	88.88	34.89	0.634	0.078	79.65	22.03	0.439	0.128	69.69	12.90	0.306	0.134
✓	✓	✗	90.17	37.08	0.642	0.081	80.67	23.15	0.415	0.126	71.48	13.72	0.296	0.116
✓	✓	✓	**92.34**	**39.39**	**0.683**	**0.061**	**83.03**	**21.76**	**0.420**	**0.118**	**74.00**	**13.76**	**0.316**	**0.097**

Table 3. **Ablation of Components.** Performance when not using Cross Attention (CroAtt), not including the Adaptive Spatial Attention module (Adapt), and replacing the 3D HOI representation with 2D image features (3DHOI). ✗ means without.

	Metric	Ours	✗CroAtt	✗Adapt	✗3DHOI
Seen	**AUC**	**92.34**	88.81	90.76	87.74
	aIOU	**39.39**	35.50	37.30	33.47
	SIM	**0.683**	0.639	0.656	0.612
	MAE	**0.061**	0.081	0.073	0.089
Unseen-Obj	**AUC**	**83.03**	76.69	79.89	75.54
	aIOU	**21.76**	20.57	20.74	20.04
	SIM	0.420	0.392	**0.422**	0.385
	MAE	**0.118**	0.142	0.121	0.139
Unseen-Aff	**AUC**	**74.00**	68.93	67.85	60.50
	aIOU	**13.76**	11.41	10.59	10.09
	SIM	**0.316**	0.266	0.274	0.240
	MAE	**0.097**	0.151	0.149	0.163

ture inputs are removed (specifically, we directly concatenate the remaining features and feed them together with the learnable query tokens into the affordance decoder), the model degrades markedly: the AUC drops by 3.82%, 5.54%, and 6.58%, with the decline most pronounced under the two *Unseen* settings. In contrast, using only the 3D HOI and object point features $\{H', P'\}$, our method matches or surpasses LASO [17] and GREAT [29]. Adding the text prompt features on top of this yields the strongest results overall, demonstrating the effectiveness of our approach at mining cross-modal evidence. **These findings answer Q2**: the proposed multimodal evidence query mechanism learns affordance knowledge from each modality and, at the same time, improves generalization in *Unseen* scenarios.

Effects of Other Components. As shown in Tab. 2, we conducted ablation studies on other components. First, we remove the cross-attention module between the textual features and the HOI features H'_o and object point clouds P'. This modification led to performance degradation across all task settings, indicating that feature fusion among different modules is beneficial. Moreover, using different textual prompts helps the model better understand both the interactive functionality and the underlying geometry. Removing the Adaptive Spatial Attention Module also leads to a clear drop in performance. This suggests that irrelevant background information in the 3D HOI space hinders the model's ability to learn geometric structures effectively.

Finally, we directly removed the 3D HOI feature and instead used a ResNet18 [7] backbone to extract features from 2D HOI images. This change caused a significant performance drop across all task scenarios, with the AUC decreas-

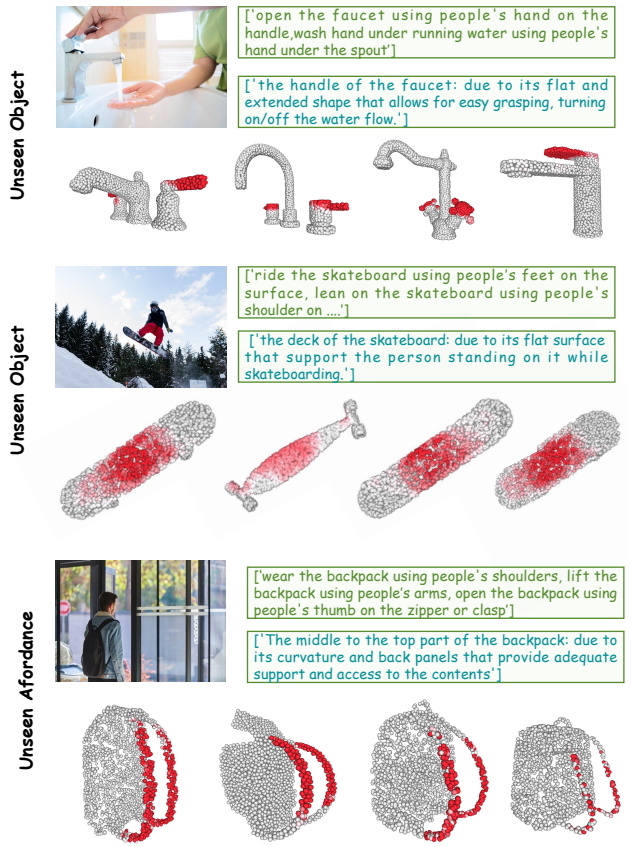

Figure 3. **Visualization Results.** We present the predicted affordances of different 3D objects, including the faucet, skateboard, and backpack, across the categories of Seen Objects, Unseen Objects, and Unseen Affordances. The depth of red indicates the affordance probability of the objects.

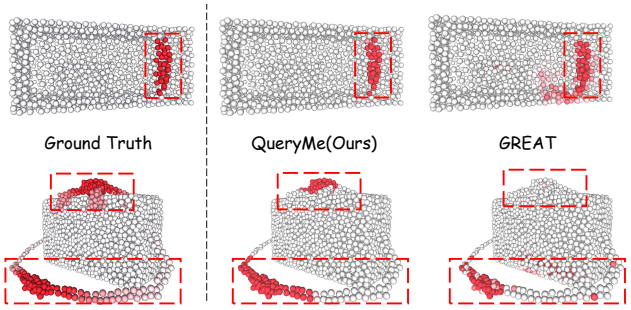

Figure 4. **Qualitative comparison** of our method, GREAT [29], and Ground Truth on the *Microwave Open* and *Bag Lift* tasks.

ing by 13.5% in the Unseen Affordance setting. Based on these observations, **we can answer Q3**: the feed-forward 3D reconstruction that maps HOI images into the 3D feature space enables the model to better learn the geometric structures underlying affordance grounding in unseen scenarios, thereby enhancing its generalization capability.

Robustness to Reconstruction Noise. In real-world 3D affordance grounding, the target object is usually obtained via reconstruction rather than a clean point cloud. To emulate this setting, we inject *spatial* noise into the object point cloud P: a proportion ρ of points is randomly selected and their coordinates are perturbed by a fixed offset δ prior to inference. We sweep ρ and report AUC at each level (Fig. 5). QueryMe exhibits the slowest degradation and retains about 69% AUC at ρ=0.85. The w/o 3D HOI feature is second best, while GREAT drops to roughly 63% under the same noise ratio. These results indicate that our multimodal evidence–driven querying and geometric reasoning in the HOI space confer strong robustness to reconstruction errors.

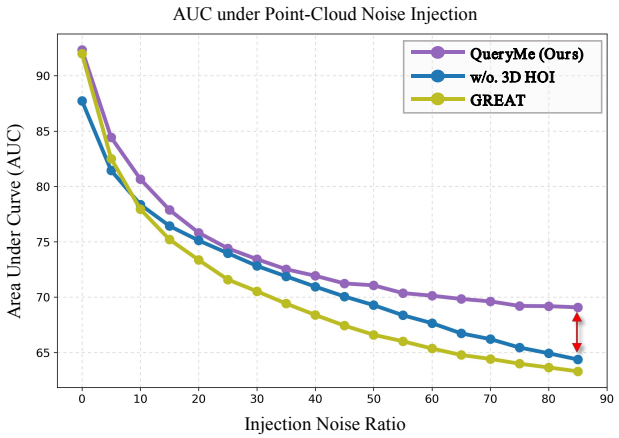

Figure 5. **Robustness to point-cloud noise.** We simulate imperfect reconstruction by injecting *spatial* perturbations into a proportion ρ of the target object's point cloud used for affordance grounding, and report AUC as ρ increases. QueryMe shows the slowest degradation and retains about 69% AUC at $\rho = 0.85$.

5. Conclusion

We present **QueryMe**, a query-driven framework for open-vocabulary 3D object affordance grounding that leverages multimodal evidence to localize functional regions. Our approach projects 2D HOI images into the 3D space, introduces an Adaptive Spatial Attention module to emphasize key interaction regions, and employs a multimodal query mechanism to jointly integrate linguistic, visual, and geometric cues for retrieving geometrically consistent functional parts. This framework enables affordance localization and geometry-based analogy reasoning, while generalizing well to unseen objects and affordances. Experimental results demonstrate that QueryMe significantly outperforms SOTA methods, particularly under unseen settings, thus validating its effectiveness and technical contributions in open-vocabulary 3D affordance understanding and grounding.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant Nos. U2574212, 62272134, 62502125 and 62402136) and the Natural Science Foundation of Shandong Province (Grant Nos. ZR2024QF064, 62502125 and ZR2025MS995).

References

[1] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. *Advances in Neural Information Processing Systems*, 35:37349–37362, 2022. 6, 7

[2] Hengshuo Chu, Xiang Deng, Qi Lv, Xiaoyang Chen, Yinchuan Li, Jianye Hao, and Liqiang Nie. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *arXiv preprint arXiv:2502.20041*, 2025. 2, 3

[3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 6

[4] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3D AffordanceNet: A Benchmark for Visual Object Affordance Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2021. 3, 6

[5] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5882–5889, 2018. 2

[6] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5880–5886. IEEE, 2023. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[8] David D. Hoffman and William A. Richards. Parts of Recognition. In *Proceedings of the Royal Society B*, pages 253–259, 1987. 2

[9] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 1

[10] Yizhou Huang, Fan Yang, Guoliang Zhu, Gen Li, Hao Shi, Yukun Zuo, Wenrui Chen, Zhiyong Li, and Kailun Yang. Resource-efficient affordance grounding with complementary depth and semantic prompts. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7788–7795. IEEE, 2025. 3

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[12] Petar Kormushev, Sylvain Calinon, and Darwin Caldwell. Imitation Learning of Positional and Force Skills Demonstrated via Kinesthetic Teaching and Human–Robot Interaction. pages 726–740, 2010. 2

[13] Di Li, Jie Feng, Jiahao Chen, Weisheng Dong, Guanbin Li, Yuhui Zheng, Mingtao Feng, and Guangming Shi. SeqAffordSplat: Scene-level Sequential Affordance Reasoning on 3D Gaussian Splatting. *arXiv preprint arXiv:2507.23772*, 2025. 2, 3

[14] Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, and Feifei Feng. Improving vision-language-action models via chain-of-affordance. *arXiv preprint arXiv:2412.20451*, 2024. 1

[15] Puhao Li, Tengyu Liu, Yuyang Li, Muzhi Han, Haoran Geng, Shu Wang, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 573–580. IEEE, 2024. 1

[16] Yu Li, Kai Cheng, Ruihai Wu, Yan Shen, Kaichen Zhou, and Hao Dong. Mobileafford: Mobile robotic manipulation through differentiable affordance learning. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*, 2024. 3

[17] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024. 1, 3, 6, 7

[18] Suhan Ling, Yian Wang, Ruihai Wu, Shiguang Wu, Yuzheng Zhuang, Tianyi Xu, Yu Li, Chang Liu, and Hao Dong. Articulated object manipulation with coarse-to-fine affordance for mitigating the effect of point cloud noise. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10895–10901. IEEE, 2024. 1

[19] Dongyue Lu, Lingdong Kong, Tianxin Huang, and Gim Hee Lee. Geal: Generalizable 3d affordance learning with cross-modal consistency. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1680–1690, 2025. 1

[20] Andrew T Miller, Steffen Knoop, Henrik I Christensen, and Peter K Allen. Automatic grasp planning using shape primitives. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, pages 1824–1829. IEEE, 2003. 3

[21] Francisco Munguia-Galeano, Satheeshkumar Veeramani, Juan David Hernández, Qingmeng Wen, and Ze Ji. Affordance-based human–robot interaction with reinforcement learning. *IEEE Access*, 11:31282–31292, 2023. 3

[22] 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. *arXiv preprint arXiv:2411.02704*, 2024. 1, 2

[23] Anh Nguyen, Dimitrios Kanoulas, Darwin G. Caldwell, and Nikos G. Tsagarakis. Object-based Affordances Detection with Convolutional Neural Networks and Dense Conditional Random Fields. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2017. 2

[24] Charles R. Qi, Li Yi, Hao Su, and et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric

Space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5, 6

[25] 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. 3

[26] Alec Radford, Jong Wook Kim, and et al. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[27] Anirban Roy and Sinisa Todorovic. A Multi-Scale CNN for Affordance Segmentation in RGB Images. In *European Conference on Computer Vision (4)*, pages 186–201, 2016. 2

[28] Mohsen Sadeghi, Hannah R Sheahan, James N Ingram, and Daniel M Wolpert. The visual geometry of a tool modulates generalization during adaptation. *Scientific Reports*, 9(1): 2731, 2019. 2

[29] Yawen Shao and et al. GREAT: Geometry-Intention Collaborative Inference for Open-Vocabulary 3D Object Affordance Grounding. *arXiv preprint arXiv:2411.19626*, 2024. 2, 3, 5, 6, 7, 8

[30] Tianmin Shu, Xiaofeng Gao, Michael S Ryoo, and Song-Chun Zhu. Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1669–1676. IEEE, 2017. 2

[31] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5, 1

[32] Josh Tobin, Lukas Biewald, Rocky Duan, Marcin Andrychowicz, Ankur Handa, Vikash Kumar, Bob McGrew, Alex Ray, Jonas Schneider, Peter Welinder, et al. Domain randomization and generative models for robotic grasping. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3482–3489. IEEE, 2018. 3

[33] Hanqing Wang, Zhenhao Zhang, Kaiyang Ji, Mingyu Liu, Wenti Yin, Yuchao Chen, Zhirui Liu, Xiangyu Zeng, Tianxiang Gui, and Hangxing Zhang. Dag: Unleash the potential of diffusion model for open-vocabulary 3d affordance grounding. *arXiv preprint arXiv:2508.01651*, 2025. 3

[34] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3, 6

[35] Yiding Wang, Bukeikhan Omarali, and Aldo A. Faisal. From gaze to action: Leveraging affordance grounding for human intention understanding. In *ICRA 2025 Workshop: Human-Centered Robot Learning in the Era of Big Data and Large Models*, 2025. 2

[36] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*, 2025. 2

[37] Zeming Wei, Junyi Lin, Yang Liu, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. 3daffordsplat: Efficient affordance reasoning with 3d gaussians. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2821–2830, 2025. 3

[38] Zeming Wei, Junyi Lin, Yang Liu, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. 3daffordsplat: Efficient affordance reasoning with 3d gaussians, 2025. 2

[39] Huadong Wu, Zhanpeng Zhang, Hui Cheng, Kai Yang, Jiaming Liu, and Ziying Guo. Learning affordance space in physical world for vision-based robotic object manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4652–4658. IEEE, 2020. 1, 3

[40] Xinli Xu, Shaocong Dong, Tingfa Xu, Lihe Ding, Jie Wang, Peng Jiang, Liqiang Song, and Jianan Li. Fusionrcnn: Lidar-camera fusion for two-stage 3d object detection. *Remote Sensing*, 15(7):1839, 2023. 1, 6, 7

[41] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10905–10915, 2023. 1, 3, 6, 7

[42] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16284–16295, 2024. 6

[43] Eiling Yee, Stacy Huffstetler, and Sharon L Thompson-Schill. Function follows form: activation of shape and function features during object identification. *Journal of Experimental Psychology: General*, 140(3):348, 2011. 2

[44] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1691–1701, 2025. 2