Diffusion Models are Open-World Affordance Learners: Leveraging Generative Priors for 3D Affordance Learning

Hanqing Wang[1,*], Zhenhao Zhang[2,*], Kaiyang Ji[2,*], Mingyu Liu[3], Wenti Yin[5], Yuchao Chen[5], Zhirui Liu[2], Xiangyu Zeng[4], Tianxiang Gui[2], Hangxing Zhang[2], Jiahao Yuan[1], Zhiqing Cui[1], Jiaxin Liu[2], Zhiyuan Ma[5,†], and Hui Xiong[1,6,†]

[1] The Hong Kong University of Science and Technology, Guangzhou, China
[2] ShanghaiTech University, Shanghai, China
[3] Zhejiang University, Hangzhou, China
[4] Nanjing University, Nanjing, China
[5] Huazhong University of Science and Technology, Wuhan, China
[6] The Hong Kong University of Science and Technology, Hong Kong, China
hwang201@connect.hkust-gz.edu.cn
* Equal contribution. † Corresponding author.

Fig. 1: Motivation: Text-to-Image Diffusion model can understand how people interact with objects. It has an awareness of affordance and can generate reasonable Human-Object Interaction (HOI) images *(Left)*. Motivated by this, we would like to find a way to transfer this rich affordance knowledge into 3D affordance grounding *(Right)*.

Abstract. 3D affordance grounding aims to understand how diverse objects can be manipulated, making it a cornerstone of embodied interaction. However, prior works struggle to generalize to out-of-distribution, open-world scenarios, leaving a critical gap between limited dataset performance and real-world application needs. Inspired by the saying: ***"What I can not create, I do not understand"***, we find generative models can generate semantically valid HOI images, which indicates inherent encoding of affordance concepts. Building on this insight, we propose DAG, the first innovative diffusion-based 3D affordance grounding framework that extracts general affordance knowledge from text-to-image diffusion models for 3D affordance prediction. Specifically, we extract the affordance priors from a diffusion model to encode HOI priors, and design

an affordance block with a multi-source affordance decoder for dense 3D affordance prediction. Extensive experiments show that DAG consistently outperforms state-of-the-art methods and exhibits strong open-world generalization, even in the challenging one-shot setting. The code of our method is released on *https://github.com/hq-King/DAG*.

Keywords: 3D vision · Affordance · Diffusion model

1 Introduction

3D affordance grounding is essential for embodied intelligence, as it unifies semantic perception with actionable interaction and serves as a fundamental bridge between perception and manipulation. While conventional object recognition focuses on identifying *what* an object is, affordance grounding addresses the more fundamental problem of understanding *how* an object can be functionally used. This capability is indispensable for intelligent agents that aim not only to perceive the environment but also to interact with it purposefully. For instance, given a natural-language query such as "Which region of the cup enables stable grasping?" or a visual demonstration of human-object interaction (HOI), a reliable affordance-grounding model should precisely identify the handle as the functional interaction region.

Existing approaches [13, 63] investigate learning affordance knowledge from HOI images and language instructions via diverse training paradigms. However, they frequently require extra supervision (e.g., bounding boxes or affordance-centric captions) and still struggle to capture consistent, generalizable regularities of real-world interactions. As a result, they only partially capture affordance semantics and do not leverage broad, transferable world knowledge, both of which are crucial for robust open-world 3D affordance grounding.

Recently, diffusion models trained on internet-scale data have revolutionized image synthesis [3, 42, 43]. By learning the distribution of real-world visual data, these models acquire strong world knowledge and can generate highly realistic and physically plausible HOI images [62]. Inspired by Feynman's insight—***"What I cannot create, I do not understand"***, we find that generative models, especially diffusion models, inherently contain affordance knowledge within their generative priors. Generation and understanding are closely coupled: a model that can reliably synthesize coherent HOI images must implicitly capture the affordances underlying human-object interactions, such as admissible manipulation modes and the spatial constraints that make them feasible.

Motivated by this, we wonder whether a large-scale text-to-image diffusion model can serve as a universal affordance learner for open-world affordance concepts, and whether its implicit knowledge can be effectively transferred to 3D affordance grounding. As visualized by the attention maps in Figure 2, diffusion models already demonstrate a strong ability to implicitly capture affordance cues embedded in HOI imagery. This motivates us to develop a novel framework that extracts and transfers rich generative priors from diffusion models to improve

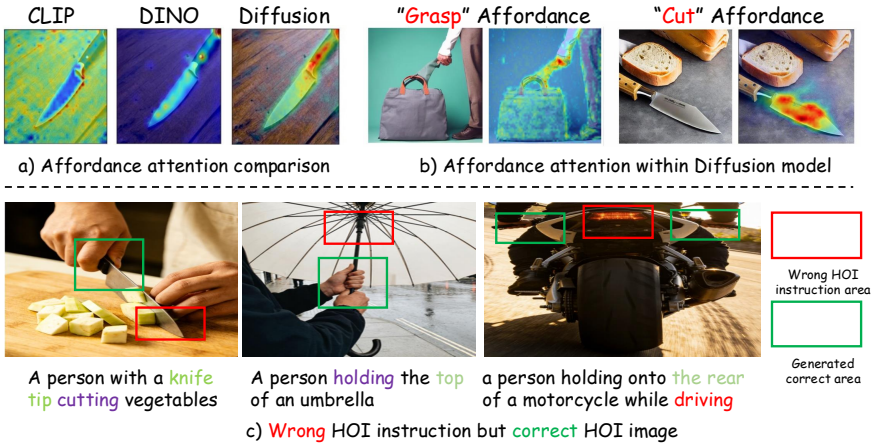

Fig. 2: We conduct some toy experiments here. As shown in a), compared with CLIP and DINO, the diffusion model can understand the affordance concept of "hold" better and locate the accurate affordance region. b) shows that the inherent affordance knowledge in the diffusion model can reveal the affordance region on both the ego and exo image. c) When the interaction area indicated by the instruction we give is incorrect, the diffusion model can still generate correct and reasonable interaction images, which indicates that it truly understands *"how to interact"* and *"where to interact"* represented by affordance internally, rather than simply mapping data distribution.

3D affordance grounding, enabling more generalizable, robust, and physically consistent functional understanding.

We propose **DAG**, the first diffusion-based 3D affordance grounding model that leverages the rich affordance knowledge encoded in diffusion models for 3D affordance grounding. An overview of our approach is illustrated in Fig. 3. At a high level, we use a pre-trained, frozen text-to-image diffusion model as a feature extractor: given an HOI image and its affordance text, we obtain multi-level internal diffusion features. Building on the extracted diffusion features, the affordance block integrates textual and visual cues to form unified affordance features. Finally, a multi-modal affordance decoder combines diffusion-derived affordance knowledge with point features extracted by a frozen point encoder to predict dense 3D affordance masks.

To summarize, our contributions are as follows:

- We verify that diffusion models inherently encode rich affordance knowledge. Based on this observation, we propose a novel framework named DAG, which unlocks the implicit affordance knowledge within diffusion models and transfer it to 3D affordance grounding.
- To better leverage affordance knowledge extracted from text-to-image diffusion models for 3D affordance grounding, we introduce an affordance block and a multi-modal affordance decoder, which effectively enhance the model's 3D grounding capability.

– We implement extensive experiments to demonstrate the effectiveness of our learning pipeline and observe noticeable gains over baselines with strong generalization capability, which highlights the effectiveness and adaptability of our approach in real-world applications.

2 Related Work

2.1 Affordance Grounding

The concept of affordance, popularized by psychologist James Gibson [14], refers to the actionable possibilities that an environment or object offers to an agent. A substantial body of work has sought to extract and model affordance knowledge from diverse data sources. In robotics, affordance learning enables robots to interact effectively with complex and dynamic real-world environments [38]. In particular, several approaches leverage affordances to relate objects, tasks, and manipulations, supporting robust robotic grasping [8, 15, 24, 29, 51]. Other lines of work learn affordances from data that can be deployed on real robots, such as human-teleoperated interaction trajectories. In computer vision, early studies primarily focused on 2D affordance detection with convolutional neural networks [22, 40, 55, 60], aiming to detect or segment action-related regions from 2D visual inputs. However, 2D affordance representations do not readily extend to the 3D spatial configurations required for physical interaction in the real world. To address this limitation, 3D AffordanceNet [9] introduced a benchmark for affordance grounding on 3D point clouds. Building on this benchmark, OpanAD [38] successfully exploits the semantic relationships between affordances by simultaneously learning the affordance text and the point feature. IAGNet [63] and MIFAG [13] pursue 3D affordance grounding from reference images. GREAT introduce affordance-centric captions to help with the affordance grounding. More recently, researchers have explored extending the reasoning capabilities of MLLMs [53, 66, 71] to affordance grounding. LMAffordance [75] applies a vision-language model to fuse 2D and 3D spatial features with semantic features. Seqafford [66] further explores how to transfer the sequential reasoning capability of large models to affordance grounding. Nevertheless, these methods typically map reference images or language instructions directly to 3D structures. As a result, they struggle to capture shared affordance regularities across diverse human–object interaction images and to model affordance knowledge in a transferable manner. In contrast, we propose leveraging diffusion models to learn more universal affordance knowledge: our observations suggest that diffusion models act as open-world affordance learners, implicitly encoding strong affordance concepts and broad world knowledge. In this paper, we investigate how to transfer and elevate this capability for 3D affordance grounding.

2.2 Image-Point Cloud Cross-Modal Learning

Multimodal learning [54, 72], which aims to integrate and exploit complementary information from heterogeneous data sources, plays a critical role in enhancing the perception and understanding capabilities of intelligent systems.

The individual limitations of each sensor type, such as LiDAR's lack of color information and cameras' inability to directly measure depth, highlight the necessity of integrating these data sources for a more complete scene understanding [64]. Image-point cloud cross-modal learning has emerged as a promising research direction to overcome the limitations of relying on either modality alone [28].To address the challenges of modality discrepancy and effective information fusion, many works have explored joint representation learning and fusion strategies [1,2,6,7,11,27,39,61,70,73]. In parallel, some studies focus on explicitly modeling the cross-modal correspondences between visual and geometric features to guide the interaction process [36, 58, 67]. In addition, recent research has also explored the use of external priors or generative models to compensate for incomplete geometry [12, 16, 21, 32, 74]. Recently, several works in affordance learning have leveraged image-point cloud cross-modal learning to achieve 3D object affordance grounding [13, 63]. These approaches align interaction cues from 2D images with 3D geometric structures, enabling the localization in 3D space. By employing cross-modal alignment and contextual modeling, such methods enhance the model's ability to perceive and reason about object affordances across modalities. However, these methods utilize CNN-based ways to directly map the connection between images and 3D structures, while our method aims at unlocking the internal affordance knowledge within the text-to-image diffusion model and transferring this into affordance grounding. Benefiting from its powerful generative prior and holistic scene understanding, the diffusion-based features offer richer semantic cues and more precise alignment with 3D geometric structures, leading to enhanced cross-modal representation and affordance reasoning.

2.3 Diffusion Models for Visual Tasks

Diffusion probabilistic models (DPMs) [10, 50] have rapidly become a leading paradigm for high-fidelity image synthesis by reversing a fixed Markovian noising process. Recent works [10,18,19,33–35,59,65,69] show that the U-Net backbones of diffusion models encode rich internal representations that can be repurposed for a wide range of downstream tasks. Learned through the denoising process, these intermediate features capture both global and local semantics, and have proven effective for image segmentation [4], depth estimation [17], and object recognition [5]. For example, [20] demonstrates that diffusion latent spaces can support image classification and related tasks without fine-tuning. Similarly, [52] exploits diffusion activations for semantic segmentation, highlighting their potential as reusable and flexible representations. [45] further apply diffusion features to cross-modal learning, such as text-to-image alignment, while [46] show improvements in self-supervised objectives (e.g., image retrieval), extending diffusion representations beyond generation. In contrast to prior efforts that primarily leverage diffusion internals for 2D vision, we are the first to systematically extract and transfer affordance knowledge from frozen diffusion representations into a 3D affordance grounding framework, enabling accurate prediction of interaction regions on point clouds even with minimal 3D supervision.

3 Method

3.1 Overview

Following previous works, our goal is to anticipate the affordance regions on the point cloud corresponding to the reference human-object interaction (HOI) image and the affordance text. Given a sample (P, I, T, y), where P is a point cloud with coordinates $P \in \mathbb{R}^{N\times 3}$, $I \in \mathbb{R}^{3\times H\times W}$ is an RGB image, T represents the affordance text, and y is the affordance category label, DAG utilizes the rich affordance knowledge within a frozen text-to-image diffusion model to assist 3D affordance grounding. Specifically, the reference image is fed into the frozen diffusion U-Net to extract the diffusion model's internal features A_v, and then we use an aggregation network to obtain the affordance priors A_g (Sec. 3.2, Sec. 3.3). With the internal features and the text features, DAG leverages the proposed Affordance Blocks to fuse them and obtain affordance tokens A_f (Sec. 3.4). The point cloud features and the [CLS] token are extracted by a pre-trained point encoder and then fed into a lightweight multi-source decoder, along with the affordance tokens A_f, to obtain the final affordance mask A_m (Sec. 3.5, Sec. 3.6). In the following sections, we describe each of these components.

3.2 Affordance Knowledge Extraction and Fusion

Advanced diffusion-based text-to-image generative models [43] use a U-Net architecture to learn the diffusion and denoising processes. The U-Net produces a number of intermediate feature maps that are large and unwieldy, varying in resolution and detail, but contain rich information about texture and semantics. At every step of the denoising process, diffusion models use the text input to infer the denoising direction of the noisy input image. This encourages visual features to correlate with rich, semantically meaningful text descriptions. Thus, the feature maps output by the U-Net blocks can be regarded as rich and dense features for affordance grounding. Our method performs a single forward pass of an input HOI image through the diffusion model to extract its visual affordance representation, as opposed to going through the entire multi-step generative diffusion process. Formally, given an input image x, we first sample a noisy image x_t at step t as:

$$x_t \triangleq \sqrt{\bar{\alpha}_t}\, x + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

where t is the diffusion time step we use, $\alpha_1, \ldots, \alpha_T$ represent a pre-defined noise schedule, and $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$, as defined in [43]. Then We utilize the proposed Affordance Semantics Encoder $\mathcal{T}$ (Sec. 3.3) to encode the affordance semantics implicitly and project the semantics into affordance knowledge A_k, and finally, we extract the text-to-image diffusion U-Net's internal features by feeding the implicit affordance semantics into the U-Net, which can be formulated as :

$$A_{v,l} = \mathrm{UNet}(x_t, A_k), \tag{2}$$

where $A_{v,l}$ is the U-Net feature of layer l. After obtaining the multi-scale features A_v from the U-Net, we propose an interpretable aggregation network that learns

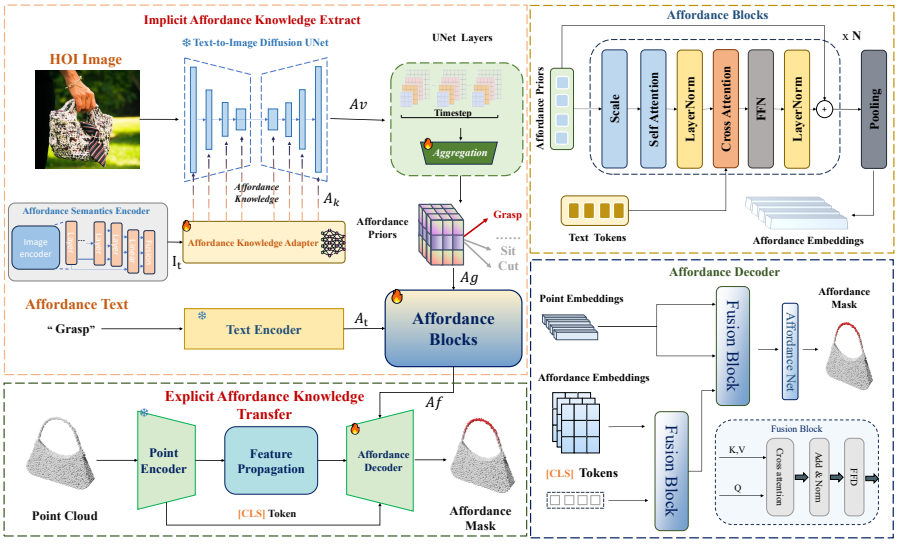

Fig. 3: DAG pipeline. Specifically, DAG employs a frozen diffusion U-Net to extract rich affordance knowledge from HOI images. An affordance semantics encoder and an implicit projector are adopted to map the extracted affordance knowledge into the U-Net structure. We then use an aggregation network to construct a unified affordance bank. Afterward, Affordance Blocks fuse the text embeddings with the affordance bank, which are further fed into a multi-modal affordance decoder. Finally, the decoder interacts with 3D point cloud embeddings to generate dense 3D affordance masks.

mixing weights across features to highlight the layers that provide the most useful information for universal affordance knowledge priors A_g, which can be formulated as:

$$A_g = \sum_{l=1}^{L} w_l \cdot A_{v,l}, \tag{3}$$

where L is the number of layers, w_l is the weight.

3.3 Affordance Semantics Encoder

To perform the single forward pass, one question is how to obtain the text embeddings for the model. Inspired by previous works [59], we adopt an Affordance Semantics Encoder (ASE) to generate an implicit text embedding from the image itself and input this embedding into the diffusion model directly. Specifically, we use a frozen CLIP image encoder to encode the image and aggregate features from the last layers. Each layer's features are first processed by a linear projection, and then all features are linearly combined via a weighted summation.

Finally, we use a learned MLP to project the implicit affordance knowledge into the U-Net. This can be described as:

$$I_t = \mathcal{T}(x), \tag{4}$$

$$A_k = \text{Proj}(I_t). \tag{5}$$

3.4 Affordance Blocks

To strengthen the generalization ability of our method, we propose an Affordance Block module to fuse the visual affordance knowledge priors A_g and the text features A_t, which are extracted by a frozen CLIP text encoder. As shown in Fig. 3, we first perform a scaling operation between A_g and A_t, and then a self-attention mechanism is used to fuse them, followed by residual connections and layer normalization. Next, we inject the implicit affordance text knowledge A_t as a condition via a cross-attention operation. Afterward, we use an FFN, residual connections, layer normalization, and an average pooling layer to obtain the affordance embeddings A_f, which can be formulated as:

$$A_r = \text{AffordanceBlock}(A_g, A_t), \tag{6}$$

$$A_f = \text{Pooling}(A_r). \tag{7}$$

3.5 Point Encoder and Propagation

We utilize a pre-trained 3D encoder to capture point cloud features. To adapt these features for dense 3D prediction, we adopt a geometry-guided upsampling and feature propagation strategy to obtain dense point features from the 3D encoder. More details about the propagation process are provided in the supplementary material.

3.6 Multi-Source Affordance Decoder

As shown in Fig. 3, we propose a lightweight multi-source affordance decoder, which utilizes the [CLS] token extracted from a frozen pre-trained 3D encoder, affordance embeddings A_f, and point features f_{p} to obtain the affordance mask A_{mask}. We first fuse the global [CLS] token and affordance embeddings A_f via cross-attention:

$$\mathbf{F}_p = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d}}\right) \cdot \mathbf{V}, \tag{8}$$

where $\mathbf{Q}$ represents the global [CLS] token and $\mathbf{K}$ and $\mathbf{V}$ represent the affordance embeddings A_f. We then apply residual connections and FFN operations and use the resulting features as $\mathbf{Q}$ to perform another fusion process with point features f_{p} as $\mathbf{K}$ and $\mathbf{V}$ to obtain the fused features $\mathbf{P}_f$. Finally, we obtain the affordance mask A_{mask} by feeding the fused features into an Affordance Net, which is an MLP:

$$A_{\text{mask}} = \text{MLP}(\mathbf{P}_f). \tag{9}$$

4 Experiment

In this section, we conduct experiments to demonstrate the efficacy of our model. We first introduce the experimental settings (Sec. 4.1). Then, we give a detailed description of our implementation (Sec. 4.2) and compare our results against the state-of-the-art (SOTA) models for 3D affordance grounding (Sec. 4.3). Lastly, we present ablation studies (Sec. 4.4) to demonstrate the effectiveness of the components and the open-world generalization ability of our method (Sec. 4.5).

4.1 Experimental Settings

Dataset. To validate the effectiveness and generalization of our method from a comprehensive perspective, we use the PIAD1 and PIAD2 datasets as our benchmarks. Both datasets are well structured and are divided into two parts: seen and unseen.

Baselines and Metrics. To provide a comprehensive and effective evaluation, we select several open-sourced cross-modal learning works [2, 23, 26, 57] and open-vocabulary affordance learning works [25, 38, 48, 63, 75]. For evaluation metrics, we follow previous works and finally choose four metrics: Area Under the Curve (AUC) [30], Mean Intersection over Union (mIoU) [41], Similarity (SIM) [49], and Mean Absolute Error (MAE) [56].

4.2 Implementation Details

Architecture. We use the stable diffusion model pre-trained on a subset of the LAION [47] dataset as our text-to-image diffusion model. Following ODISE [59], we extract feature maps from every three of its U-Net blocks and resize them to create a feature pyramid. Following [59], we set the time step used for the diffusion process to $t = 0$. By default, we use the clip text encoder to encode the affordance text. As for point cloud, we use pre-trained Uni3D [68] as our point cloud encoder. During training, we use the Adam [31] optimizer with a learning rate set to 1e-4.

Training objectives. Our model seeks to transfer the affordance knowledge in the text-to-diffusion model into a 3D affordance decoder. Thus, we solely employ Dice loss [37] and Binary Cross Entropy(BCE) loss [44] to guide the segmentation mask prediction:

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{Dice}. \tag{10}$$

Following previous research, we show the comparison results on PIAD [63]. We split the dataset into two parts, including *Seen* and *Unseen*. *Seen*: This default setting maintains similar distributions of object classes and affordance types across both training and testing phases. *Unseen*: This configuration is specifically designed to evaluate the model's ability to generalize affordance knowledge.

Table 1: Main Results. The overall results of all comparative methods, the best results are in **bold** and the second results are in underline.

	Method	$mIoU\uparrow$	$AUC\uparrow$	$SIM\uparrow$	$MAE\downarrow$
Seen	ReferTrans	11.32	78.89	0.497	0.129
	ReLA	12.08	79.13	0.483	0.125
	PFU	12.31	77.50	0.432	0.135
	XMF	12.94	78.24	0.441	0.127
	IAGNet	20.51	84.85	0.545	0.098
	LASO	21.14	86.12	0.559	0.092
	GREAT	22.72	87.94	0.594	0.087
	Ours	**24.84±0.5**	**90.16±0.3**	**0.637±0.03**	**0.078±0.01**
Unseen	ReferTrans	7.130	67.40	0.327	0.151
	ReLA	7.320	68.20	0.323	0.147
	PFU	5.330	61.87	0.330	0.193
	XMF	5.680	62.58	0.342	0.188
	IAGNet	7.950	71.84	0.352	0.127
	LASO	8.110	71.98	0.366	0.126
	GREAT	8.820	73.61	0.384	0.124
	Ours	**9.730±0.4**	**76.69±0.3**	**0.414±0.05**	**0.120±0.02**

In this setup, certain affordance-object pairings are deliberately omitted from the training set but introduced during testing. The training is done on one A100 GPU for 80 epochs for the main experiments. And we show the effectiveness of our model by answering the following questions: **Q1: How is our model compared to other baselines? Q2: How effective are the proposed components?**

4.3 Main Results(Q1)

In Table 1, our model demonstrates superior performance across all evaluation metrics compared to the baseline:

DAG vs.Other methods. To evaluate the generalization capability of our method, we analyze its performance across seen and unseen settings. Our proposed method (DAG) demonstrates superior performance compared to previous methods on both seen and unseen settings. This consistent superiority across datasets validates the robustness and effectiveness of DAG in the open world.

As can be seen in Fig. 4, we present a qualitative comparison of affordance predictions under both seen and unseen settings. Our method, DAG, shows strong generalization, extracting the general affordance knowledge within

Table 2: Main Results. The overall results of all comparative methods on the PIAD2 dataset, the best results are in **bold** and the second results are in underline. DAG performs well even on the unseen affordance setting.

	Method	$mIoU\uparrow$	$AUC\uparrow$	$SIM\uparrow$	$MAE\downarrow$
Seen	FRCNN	33.55	87.05	0.600	0.082
	XMF	33.91	87.39	0.604	0.078
	IAGNet	34.29	89.03	0.623	0.076
	LASO	34.88	90.34	0.627	0.077
	GREAT	38.03	91.99	0.676	0.067
	Ours	**47.19±0.3**	**94.71±0.4**	**0.779±0.02**	**0.062±0.02**
Unseen Obj	FRCNN	18.08	72.20	0.362	0.152
	XMF	17.40	74.61	0.361	0.126
	IAGNet	16.78	73.03	0.351	0.123
	LASO	16.05	73.32	0.354	0.123
	GREAT	20.16	79.57	0.402	0.109
	Ours	**28.93±0.7**	**85.41±0.3**	**0.592±0.05**	**0.102±0.02**
Unseen Aff	FRCNN	7.88	58.09	0.208	0.160
	XMF	7.96	59.08	0.210	0.156
	IAGNet	8.11	60.99	0.225	0.152
	LASO	8.37	64.07	0.228	0.140
	GREAT	12.05	69.81	0.290	0.127
	Ours	**16.09±0.4**	**75.23±0.3**	**0.372±0.05**	**0.123±0.02**

human-object images. And DAG can also maintain consistent and robust performance across diverse scenarios.

4.4 Ablation Study and Analysis(Q2)

In this section, we conduct a comprehensive ablation study and analysis to investigate the effect of different framework designs.

Effectiveness of Affordance Blocks. Table. 4 reports the impact on evaluation metrics of the Affordance Blocks. Introducing this module results in a substantial improvement over the baseline, which underscores that our method enables deep integration of high-dimensional semantic features representing instructions with dense features from point clouds and rich affordance world knowledge within the diffusion model, thereby making affordance reasoning more effective.

Effectiveness of [CLS] Token. The influence of **[CLS] Token** on evaluation metrics is also shown in Table. 4. With the integration of the **[CLS]**-guided mask,

Table 3: Ablation study on the Text Caption. *Verb* represents the affordance verb text. The best results are in **bold** and the second results are in underline

Variants	$mIoU\uparrow$	$AUC\uparrow$	$SIM\uparrow$	$MAE\downarrow$
Empty	13.9	80.5	0.462	0.116
BLIP	18.4	85.5	0.542	0.108
Verb	20.2	86.7	0.578	0.094
Ours(ASE)	**24.84±0.5**	**90.16±0.3**	**0.637±0.03**	**0.078±0.01**

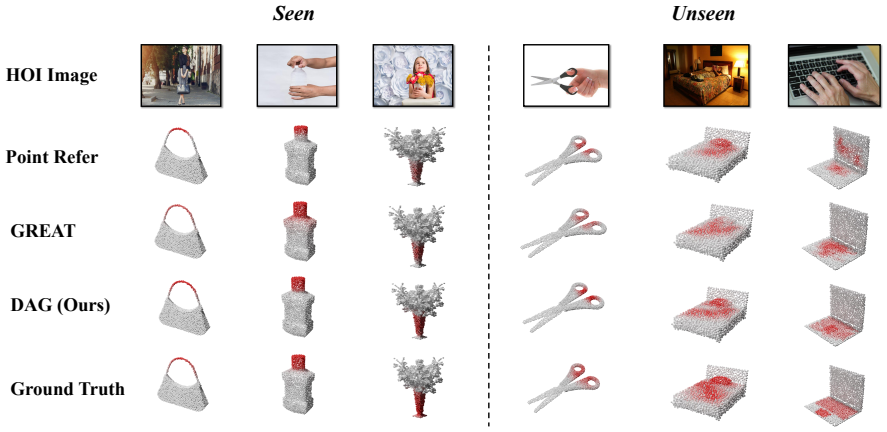

Fig. 4: Affordance Visualization. DAG achieves more accurate results in both seen and unseen settings. For more visualization results, please check our Appendix.

results of both settings can be improved, suggesting that the **[CLS]** token can introduce the global point features, which are essential for affordance grounding.

Different Captioner Strategy. As shown in the Table. 3, we construct several baselines to show the effectiveness of our implicit captioning module: Affordance Semantics Encoder. Since our implicit self-prompt captioning module derives its caption from a text-image discriminative model trained on Internet-scale data, it is able to generalize best among all variants compared.

Different Affordance Extractor. We conducted comparative experimental analysis on the superiority of using text-to-image diffusion as the affordance extractor. We selected commonly used image feature networks (ResNet), text-image pair pre-trained model (CLIP), and visual self-supervised model (DINOv2) to compare with diffusion. As shown in Table 5, the diffusion affordance extractor outperforms all the variants. As the saying goes, ***What I can not create, I do not understand***, the diffusion model can understand both the global semantic concept and the details of the image, which comes from its generation training

Table 4: We investigate the improvement of Affordance Block and [CLS] Token on the model performance based on the baseline. The best results are in **bold** and the second results are in underline

	Affordance Block	[CLS]	*mIOU*↑	*AUC*↑	*SIM*↑	*MAE*↓
Seen	×	×	10.02	79.95	0.398	0.124
	✓	×	18.56	87.04	0.582	0.111
	×	✓	12.97	84.52	0.470	0.116
	✓	✓	**24.84±0.5**	**90.16±0.3**	**0.637±0.03**	**0.078±0.01**
Unseen	×	×	8.45	67.84	0.323	0.138
	✓	×	9.02	74.30	0.374	0.126
	×	✓	8.63	70.51	0.360	0.131
	✓	✓	**9.730±0.4**	**76.69±0.7**	**0.414±0.05**	**0.120±0.02**

Table 5: Ablation study on Affordance Extractor. We select several representative models to conduct experiments to compare the affordance knowledge extraction ability. DAG (Diffusion-based) outperforms on all metrics.

Variants	*mIoU*↑	*AUC*↑	*SIM*↑	*MAE*↓
ResNet	13.7	82.3	0.514	0.128
CLIP	21.2	86.2	0.573	0.091
DINOv2	22.3	87.5	0.582	0.084
Ours	**24.84±0.5**	**90.16±0.3**	**0.637±0.03**	**0.078±0.01**

process. While other models lack semantic concepts (such as ResNet, DINO) or only understand the global information of the image, but lose details (CLIP).

Timestep sensitivity. In our default setting, we empirically set the diffusion timestep $t = 0$ to acquire high-quality feature representations. To fully validate the rationality and superiority of this key hyperparameter choice, we conduct a comprehensive ablation study on the PIAD1 seen benchmark. For a fair comparison, all experimental configurations are strictly consistent across different timesteps. As quantitatively illustrated in Table 6, the model achieves the optimal performance on all evaluation metrics when the timestep is set to 0. In contrast, larger diffusion timesteps will introduce excessive noise interference in the feature extraction process, which blurs detailed texture information, weakens the dense spatial localization capability of the model, and ultimately reduces the accuracy and robustness of prediction results. The above ablation results

Table 6: Timestep ablation.

t	AUC↑	mIoU↑	SIM↑	MAE↓
0	**90.16**	**24.85**	**0.637**	**0.078**
50	88.92	23.11	0.619	0.082
100	88.11	21.93	0.608	0.084

firmly demonstrate that setting $t = 0$ is the most suitable configuration for our diffusion-based feature extraction module.

4.5 More Analysis about the Open-World Generalization Ability

Performance on Open-World settings. To further validate the generalization ability of DAG in open-world scenarios, we conduct extensive experiments under the few-shot setting, where the model is trained with very little data, and is expected to recognize unseen objects and affordances during inference. Specifically, we evaluate the model when only a small number of labeled examples are available for each affordance category. As demonstrated in the Table 7, even under the challenging 1-shot setting, our model still maintains competitive affordance grounding performance, which clearly demonstrates its powerful ability to generalize to novel categories with limited supervision. As the number of shots increases from 1 to 5, the performance of DAG gradually rises and steadily outperforms competing approaches, which are trained with the full dataset, further verifying the effectiveness and robustness of unlocking affordance knowledge from diffusion models for 3D affordance grounding. And we also provid with some few shots experiments results of other methods in the supplementary material.

Visualization Results on Partial Point Clouds. In the real physical world, due to the limited viewing angles observed by embodied agents and the limitations of sensor perception, the point clouds they obtain are partially incomplete and with noise. This poses a huge challenge for the practical application of algorithms. We selected incomplete point clouds from the 3D AfferdanceNet dataset to test the algorithm. As shown in Figure 5, even if the point clouds are partial, DAG can roughly predict the affordance region, demonstrating the powerful generalization applicability.

Fig. 5: Visualization Results on Partial Point Clouds. Even if the input point cloud is incomplete, DAG can still predict the affordance area well, which shows the strong generalization ability of our method.

Table 7: Performance of our DAG under different few-shot settings. DAG can achieve robust and competitive performance on limited and Out-of-Distribution Data, which shows that our model can truly understand the concept of affordance.

Setting	*mIoU*↑	*AUC*↑	*SIM*↑	*MAE*↓
Ours (5 shots) (2% training data)	15.74	82.04	0.4869	0.121
Ours (3 shots) (1% training data)	14.34	80.22	0.4955	0.124
Ours (1 shot) (less than 1%)	12.98	79.42	0.4952	0.125
PFU (All data)	12.31	77.50	0.432	0.135
IAGNet (All data)	20.51	84.85	0.545	0.098
Ours (All data)	**24.84±0.5**	**90.16±0.3**	**0.637±0.03**	**0.078±0.01**

5 Conclusion and Future Work

We present a novel framework, **DAG**, which is designed to unlock the rich affordance knowledge within frozen text-to-image diffusion models for 3D affordance grounding. By leveraging the frozen U-Net to extract affordance priors in a feature pyramid manner and integrating Affordance Blocks, an Affordance Semantics Encoder, and a multi-source Affordance Decoder, DAG achieves precise affordance prediction, and extensive experiments demonstrate that DAG outperforms previous state-of-the-art methods and show strong open-world generalization ability. Future work will focus on transferring rich affordance knowledge to 3D grounding despite limited 3D data.

6 Acknowledgement

This paper is supported by the National Natural Science Foundation of China (No. 62406161).

References

1. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9902–9912 (2022)
2. Aiello, E., Valsesia, D., Magli, E.: Cross-modal learning for image-guided point cloud shape completion. Advances in Neural Information Processing Systems **35**, 37349–37362 (2022)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., et al.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
5. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19830–19843 (2023)

6. Chen, Z., Xu, H., Chen, W., Zhou, Z., Xiao, H., Sun, B., Xie, X., et al.: Pointdc: Unsupervised semantic segmentation of 3d point clouds via cross-modal distillation and super-voxel clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14290–14299 (2023)
7. Choo, Y.S., Kim, B., Kim, H.S., Park, Y.S.: Supervised contrastive learning for 3d cross-modal retrieval. Applied Sciences **14**(22), 10322 (2024)
8. Chu, H., Deng, X., Lv, Q., Chen, X., Li, Y., Hao, J., Nie, L.: 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. arXiv preprint arXiv:2502.20041 (2025)
9. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1778–1787 (2021)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
11. Dong, R., Qi, Z., Zhang, L., Zhang, J., Sun, J., Ge, Z., Yi, L., Ma, K.: Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=8Oun8ZUVe8N
12. Du, Z., Dou, J., Liu, Z., Wei, J., Wang, G., Xie, N., Yang, Y.: Cdpnet: cross-modal dual phases network for point cloud completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1635–1643 (2024)
13. Gao, X., Zhang, P., Qu, D., Wang, D., Wang, Z., Ding, Y., Zhao, B.: Learning 2d invariant affordance knowledge for 3d affordance grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3095–3103 (2025)
14. Gibson, J.J.: The theory of affordances. Hilldale, USA **1**(2), 67–82 (1977)
15. Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. arXiv preprint arXiv:2502.21257 (2025)
16. Kasten, Y., Rahamim, O., Chechik, G.: Point cloud completion with pretrained text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 12171–12191 (2023)
17. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9492–9502 (2024)
18. Kim, H., Baik, S., Joo, H.: David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10330–10341 (2025)
19. Kim, H., Han, S., Kwon, P., Joo, H.: Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In: European Conference on Computer Vision. pp. 400–419. Springer (2024)
20. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2206–2217 (2023)
21. Li, A., Zhu, Z., Wei, M.: Genpc: Zero-shot point cloud completion via 3d generative priors. arXiv preprint arXiv:2502.19896 (2025)
22. Li, G., Sun, D., Sevilla-Lara, L., Jampani, V.: One-shot open affordance learning with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3086–3096 (2024)

23. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. Advances in neural information processing systems **34**, 19652–19664 (2021)
24. Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J., Dong, H.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18061–18070 (2024)
25. Li, Y., Zhao, N., Xiao, J., Feng, C., Wang, X., Chua, T.s.: Laso: Language-guided affordance segmentation on 3d object. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14251–14260 (2024). https://doi.org/10.1109/CVPR52733.2024.01351
26. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23592–23601 (2023)
27. Liu, G., Li, H., Gao, R., Li, X., Ma, Z., Fang, T.: Dream3davatar: Text-controlled 3d avatar reconstruction from a single image (2025), https://arxiv.org/abs/2509.13013
28. Liu, H., Duan, T.: Cross-modal collaboration and robust feature classifier for open-vocabulary 3d object detection. Sensors **25**(2), 553 (2025)
29. Liu, L., Han, Y., Yi, P., Yu, W., Wang, H., Du, H., Yuan, E., Yuan, Z., Feng, R., Liu, M., et al.: Affordance2action: Task-conditioned scene-level affordance grounding for real-time manipulation. arXiv preprint arXiv:2606.04172 (2026)
30. Lobo, J.M., Jiménez-Valverde, A., Real, R.: Auc: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography **17**(2), 145–151 (2008)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
32. Ma, C., Yang, Y., Guo, J., Pan, F., Wang, C., Guo, Y.: Unsupervised point cloud completion and segmentation by generative adversarial autoencoding network. Advances in Neural Information Processing Systems **35**, 3556–3568 (2022)
33. Ma, Z., zhihuan yu, Li, J., Zhou, B.: Lmd: Faster image reconstruction with latent masking diffusion (2023), https://arxiv.org/abs/2312.07971
34. Ma, Z., Zhang, Y., Jia, G., Zhao, L., Ma, Y., Ma, M., Liu, G., Zhang, K., Li, J., Zhou, B.: Efficient diffusion models: A comprehensive survey from principles to practices (2024), https://arxiv.org/abs/2410.11795
35. Ma, Z., Zhao, L., Qi, B., Zhou, B.: Neural residual diffusion models for deep scalable vision generation (2024), https://arxiv.org/abs/2406.13215
36. Mao, A., Tang, Y., Huang, J., He, Y.: Dmf-net: Image-guided point cloud completion with dual-channel modality fusion and shape-aware upsampling transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6063–6071 (2025)
37. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
38. Nguyen, T., Vu, M.N., Vuong, A., Nguyen, D., Vo, T., Le, N., Nguyen, A.: Open-vocabulary affordance detection in 3d point clouds. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5692–5698. IEEE (2023)
39. Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L.: Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In: International Conference on Machine Learning. pp. 28223–28243. PMLR (2023)

40. 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)
41. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: International symposium on visual computing. pp. 234–244. Springer (2016)
42. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Ruby, U., Yendapalli, V.: Binary cross entropy with deep learning technique for image classification. International Journal of Advanced Trends in Computer Science and Engineering **9** (10 2020). https://doi.org/10.30534/ijatcse/2020/175942020
45. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
46. Samuel, D., Ben-Ari, R., Levy, M., Darshan, N., Chechik, G.: Where's waldo: Diffusion features for personalized segmentation and retrieval. Advances in Neural Information Processing Systems **37**, 128160–128181 (2024)
47. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
48. Shao, Y., Zhai, W., Yang, Y., Luo, H., Cao, Y., Zha, Z.J.: Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. arXiv preprint arXiv:2411.19626 (2024)
49. Swain, M.J., Ballard, D.H.: Color indexing. International journal of computer vision **7**(1), 11–32 (1991)
50. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36**, 1363–1389 (2023)
51. Tang, Y., Zhang, S., Hao, X., Wang, P., Wu, J., Wang, Z., Zhang, S.: Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. arXiv preprint arXiv:2503.00778 (2025)
52. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3554–3563 (2024)
53. Wang, H., Liu, M., Chen, X., Ma, C., Zhong, Y., Yin, W., Liu, Y., Cui, Z., Yuan, J., Dai, L., et al.: Videoafford: Grounding 3d affordance from human-object-interaction videos via multimodal large language model. arXiv preprint arXiv:2602.09638 (2026)
54. Wang, H., Tian, Y., Liu, M., Zhang, Z., Zhu, X.: Sdeval: Safety dynamic evaluation for multimodal large language models (2026), https://arxiv.org/abs/2508.06142
55. 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)
56. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate research **30**(1), 79–82 (2005)
57. Xu, D., Anguelov, D., Jain, A.: Pointfusion: Deep sensor fusion for 3d bounding box estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 244–253 (2018)
58. Xu, H., Long, C., Zhang, W., Liu, Y., Cao, Z., Dong, Z., Yang, B.: Explicitly guided information interaction network for cross-modal point cloud completion. In: European Conference on Computer Vision. pp. 414–432. Springer (2024)
59. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
60. Xu, P., Yadong, M.: Weakly-supervised affordance grounding guided by part-level semantic priors. In: The Thirteenth International Conference on Learning Representations
61. Yan, X., Zhan, H., Zheng, C., Gao, J., Zhang, R., Cui, S., Li, Z.: Let images give you more: Point cloud cross-modal training for shape analysis. Advances in Neural Information Processing Systems **35**, 32398–32411 (2022)
62. Yang, J., Li, B., Yang, F., Zeng, A., Zhang, L., Zhang, R.: Boosting human-object interaction detection with text-to-image diffusion model. arXiv preprint arXiv:2305.12252 (2023)
63. Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., Zha, Z.J.: Grounding 3d object affordance from 2d interactions in images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10905–10915 (2023)
64. Yao, G., Xuan, Y., Li, X., Pan, Y.: Cmr-agent: Learning a cross-modal agent for iterative image-to-point cloud registration. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13458–13465. IEEE (2024)
65. 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)
66. Yu, C., Wang, H., Shi, Y., Luo, H., Yang, S., Yu, J., Wang, J.: Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1691–1701 (2025)
67. Yuan, M., Fu, K., Li, Z., Meng, Y., Wang, M.: Pointmbf: A multi-scale bidirectional fusion network for unsupervised rgb-d point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17694–17705 (2023)
68. Zhang, B., Yuan, J., Shi, B., Chen, T., Li, Y., Qiao, Y.: Uni3d: A unified baseline for multi-dataset 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9253–9262 (2023)
69. Zhang, H., Carlone, L.: H2oflow: Grounding human-object affordances with 3d generative models and dense diffused flows. arXiv preprint arXiv:2510.21769 (2025)
70. Zhang, Z., Liu, J., Shi, Y., Wang, J.: Unihm: Unified dexterous hand manipulation with vision language model (2026), https://arxiv.org/abs/2603.00732

71. Zhang, Z., Shi, Y., Yang, L., Ni, S., Ye, Q., Wang, J.: Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. arXiv preprint arXiv:2505.18947 (2025)
72. Zhang, Z., Wang, H., Zeng, X., Cheng, Z., Liu, J., Yan, H., Liu, Z., Ji, K., Gui, T., Hu, K., Chen, K., Fan, Y., Pan, M.: Hoid-r1: Reinforcement learning for open-world human-object interaction detection reasoning with multimodal large language model (2025), https://arxiv.org/abs/2508.11350
73. Zhou, H., Peng, X., Luo, Y., Wu, Z.: Pointcmc: cross-modal multi-scale correspondences learning for point cloud understanding. Multimedia Systems **30**(3), 138 (2024)
74. Zhou, J., Song, W., Wang, M., Tan, H., Li, N., Liu, X.: Fine-grained text and image guided point cloud completion with clip model. Neurocomputing **631**, 129768 (2025)
75. Zhu, H., Kong, Q., Xu, K., Xia, X., Deng, B., Ye, J., Xiong, R., Wang, Y.: Grounding 3d object affordance with language instructions, visual observations and interactions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17337–17346 (2025)