Preprints.org

Review

Not peer-reviewed version

From Passive Perception to Active Interaction: A Survey of Affordance Learning for Embodied AI

Gen Li [‡,†], Hanqing Wang [†], Jingliang Li [†], Yifan Han [†], Jindou Jia, Tao Lin, Yuanzhe Liu, Yutong Wang, Bo Zhao, Fangqiang Ding, Anh Nguyen, Laura Sevilla-Lara, Huazhe Xu, Gregory S. Chirikjian, Marc Pollefeys, Oier Mees, Hui Xiong [*], Jianfei Yang [*]

Posted Date: 7 October 2026

Keywords: affordance learning; embodied AI; computer vision; robot learning

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

From Passive Perception to Active Interaction: A Survey of Affordance Learning for Embodied AI

Gen Li [1,†,‡], Hanqing Wang [2,†], Jingliang Li [1,†], Yifan Han [3,†], Jindou Jia [1], Tao Lin [3], Yuanzhe Liu [4], Yutong Wang [5], Bo Zhao [3], Fangqiang Ding [2], Anh Nguyen [6], Laura Sevilla-Lara [7], Huazhe Xu [8], Gregory S. Chirikjian [9], Marc Pollefeys [10], Oier Mees [10], Hui Xiong [2,*] and Jianfei Yang [1,*]

[1] Nanyang Technological University
[2] HKUST(GZ)
[3] Shanghai Jiao Tong University
[4] UIUC
[5] The University of Sydney
[6] University of Liverpool
[7] The University of Edinburgh
[8] Tsinghua University
[9] MBZUAI
[10] JETH Zurich
* Correspondence: jianfei.yang@ntu.edu.sg (J.Y.); xionghui@ust.hk (H.X.)
† Equal Contribution
‡ Project Lead

Abstract

Embodied AI aims to develop agents that perceive, reason, and act in physical environments. Beyond recognizing objects and scenes, agents must identify possible interactions, determine their relevance to the current task and embodiment, and translate them into executable behavior. Affordance learning supports this process by grounding perception in action possibilities and connecting it with task reasoning and control. Existing affordance surveys, however, provide only partial coverage of the field, often focusing on isolated aspects of perception or robotic interaction and lacking a comprehensive view of the full perception–reasoning–action pipeline in embodied intelligence. To bridge this gap, this survey presents a unified framework organized around three complementary roles: affordance perception identifies possible interactions and their locations; affordance reasoning selects opportunities relevant to the task, context, and agent state; and affordance-guided action uses affordance representations for planning and policy learning. Within this framework, we develop fine-grained taxonomies, compare representative methods, and trace the field's progression from passive visual recognition toward action-aware embodied learning. We further organize representative datasets and evaluation protocols and identify open challenges and future directions. Overall, this survey clarifies how affordance learning connects perception, reasoning, and action to support generalizable and physically grounded embodied intelligence.

Keywords: affordance learning; embodied AI; computer vision; robot learning

1. Introduction

Embodied intelligence studies agents that perceive, reason, and act in physical environments. Unlike image classifiers or passive scene understanding systems, embodied agents must use perception to support interaction: they need to identify task-relevant objects and actionable parts, anticipate the effects of actions, and assess whether those effects are achievable with their bodies and skills. This makes action-oriented perception a central requirement for embodied AI.

Affordance provides a way to ground perception in possible interactions. While semantic categories describe what an object is, affordances characterize what an agent can do with it. Recognizing a

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

mug, drawer, or chair is only a first step: an agent must further infer how the mug can be grasped or used for pouring, where the drawer should be manipulated, and whether the chair should serve as a seat, provide support, or be treated as an obstacle. Such action possibilities are not fixed object properties but arise from relations among objects, tasks, environments, and embodiment. For example, a cup may be grasped by its handle or body depending on the intended action and end-effector; an occupied tabletop may no longer afford placement; and a tool may become useful only when paired with an appropriate target.

In embodied AI, affordance learning infers how visual, geometric, semantic, and physical cues enable interaction with the environment. Beyond recognizing objects or localizing regions, it identifies which elements are actionable, what interactions they support, why they are relevant to a task, and how they constrain feasible behavior. Affordance therefore provides a functional intermediate representation connecting perception, task-level reasoning, and executable control. Early learning-based work largely treated affordance learning as visual prediction, recognizing functional categories or localizing action regions in RGB observations using labels or masks [1–4]. Although effective for identifying visible functional regions, these formulations were often tied to predefined labels and offered limited guidance for physical execution. With the rise of embodied AI, the field has expanded toward task-oriented and action-aware understanding. Recent methods ground affordances in 3D space [5,6], derive contact and motion cues from human interactions and egocentric videos [7–10], use language and multimodal models to identify task-relevant objects and parts [11–13], and incorporate affordances into robot planning and policy learning [14–18]. Collectively, these developments shift affordance learning from appearance-driven recognition toward interactions that are contextualized by task goals and physically executable by embodied agents.

Several surveys have examined different aspects of the affordance literature. Visual affordance surveys examine function understanding, recognition, segmentation, and reproducibility [19–21], while robotic reviews cover cognitive foundations, developmental learning, affordance relations, and reinforcement-learning formulations [22–25]. These works provide valuable foundations, yet important gaps remain in the context of modern embodied AI. Perception-centered surveys provide limited coverage of how affordances are selected and interpreted under language instructions, implicit intents, tool-use reasoning, and feasibility-aware planning [11,13,26,27]. Conversely, robotics-centered reviews do not fully capture recent advances driven by foundation models, large-scale robot learning, vision-language-action models, and video-action models [28–35]. What remains missing is a unified view of affordance learning that explains how these threads jointly connect perception, reasoning, and action in support of embodied intelligence.

To address this gap, this survey examines affordance learning from a comprehensive embodied intelligence perspective. As illustrated in Figure 1, we organize the literature around three core roles: **affordance perception**, **affordance reasoning**, and **affordance-guided action**. Together, they describe how an embodied agent identifies actionable opportunities, reasons about their task relevance, and turns them into executable behavior. This organization also defines the survey scope: we include studies in which affordance plays an explicit role in problem formulation, supervision, model output, planning process, or evaluation. General embodied AI methods are considered only when they explicitly ground, utilize, or evaluate action possibilities. This criterion accommodates recent developments in large language model (LLM)- and multimodal large language model (MLLM)-based reasoning, agentic systems, and vision-language-action models while maintaining the focus on affordance learning.

We first introduce the background in Section 2, covering definition, unified problem formulation, and evolution of affordance learning. Our technical review then begins with affordance perception, which asks what is actionable and where it is located. Section 3 covers visual, spatial, interaction-driven, and generalizable affordance grounding at both object and scene levels. Second, affordance reasoning asks why and when an affordance is relevant. Section 4 examines how task-relevant objects, parts, and interactions are selected through relation-centric modeling, the interpretation of language

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

instructions or implicit intentions, and agentic workflows. Third, affordance-guided action asks how affordances are translated into executable behavior. Section 5 reviews hierarchical affordance-to-action pipelines and affordance-integrated policy learning, in which affordance provides functional structure for action generation. The three roles form a non-exclusive framework: each method is discussed according to its primary contribution, while cross-role methods may be revisited from complementary perspectives. Finally, Section 6 summarizes datasets, benchmarks, and evaluation protocols, while Section 7 discusses open challenges and future directions.

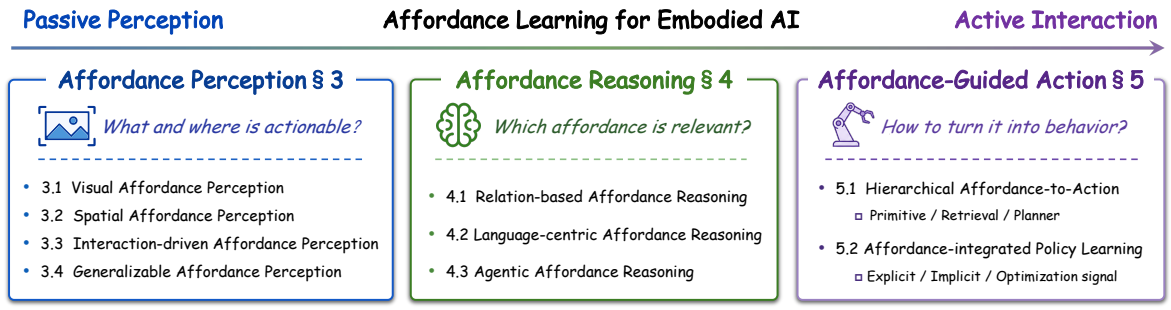


Figure 1. Taxonomy of affordance learning in embodied AI across perception, reasoning, and affordance-guided action.

In summary, this survey makes the following contributions:

- We present a unified framework for affordance learning in embodied AI, organized around perception, reasoning, and affordance-guided action. It clarifies how actionable opportunities are grounded, contextualized, and translated into executable behavior.
- Within this framework, we develop fine-grained taxonomies and compare representative methods along consistent dimensions of inputs, representations, supervision, generalization, and execution.
- We organize representative datasets and evaluation protocols, and outline open challenges and future directions, providing the community with a structured resource for future research in embodied affordance learning.

2. Background

2.1. Definition and Problem Formulation

The term affordance, introduced by Gibson [36], refers to action possibilities that the environment offers to an agent. Early research on active perception and interaction laid foundations for understanding object functionality [37,38]. Subsequent approaches learned tool affordances [39] and visual cues for predicting possible interactions [40] and recognizing functional object classes [41]. Building on these efforts, affordance learning has expanded across computer vision, robotics, and embodied AI, encompassing tasks such as affordance segmentation, contact prediction, action-region grounding, and affordance-conditioned robot control [2,4,10,42–44].

Different problems vary in supervision, input modalities, output representations, and downstream uses. To provide a common notation, we formulate affordance learning as the problem of estimating structured affordance information from observations, task information, and auxiliary evidence. Let o denote an observation, such as an RGB or RGB-D image, video, or 3D representation; l, language or task information, such as an affordance word, action query, or natural-language instruction; and c, contextual evidence, such as human–object interactions, demonstrations, support examples, or retrieved memories. An affordance model f_θ can then be formulated as

$$\mathcal{F} = f_\theta(o, l, c) = (a, r, \eta), \tag{1}$$

where the structured affordance representation $\mathcal{F}$ contains three complementary components. The action or affordance label a describes **what** action is possible, such as grasping, opening, or pulling; the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

grounded region r specifies **where** the action can be performed, such as an object part, mask, heatmap, keypoint, contact point, or 3D spatial location; and the execution-related specification η captures **how** the affordance can be used for embodied action, such as a 6-DoF pose, action score, motion direction, or post-contact trajectory. In practice, inputs l and c are optional: purely visual methods may use only o, language-conditioned methods use (o, l), and transfer-based methods further condition on c. Depending on the problem setting, some output components may also be omitted or left implicit. For example, visual affordance segmentation mainly predicts affordance labels a along with their grounded regions r, while robot manipulation methods often further model η to enable action execution.

This formulation connects the main settings reviewed in this survey. Visual and spatial perception primarily focus on grounding relevant regions r, while interaction-driven methods leverage c to transfer knowledge from human interactions or demonstrations [7,42,45]. Affordance reasoning incorporates task information l to identify task-relevant objects, parts, or relationships [11,13,46], and affordance-guided action supplies $\mathcal{F}$ to a planner, policy, or controller [15,16,47].

2.2. Evolution of Affordance Learning Paradigms

Affordance learning has evolved from supervised visual prediction of action regions toward action-grounded embodied learning. Figure 2 traces representative advances over the past decade, while the following discussion summarizes five interrelated paradigm shifts in supervision, grounding space, output representation, generalization, and integration with action.

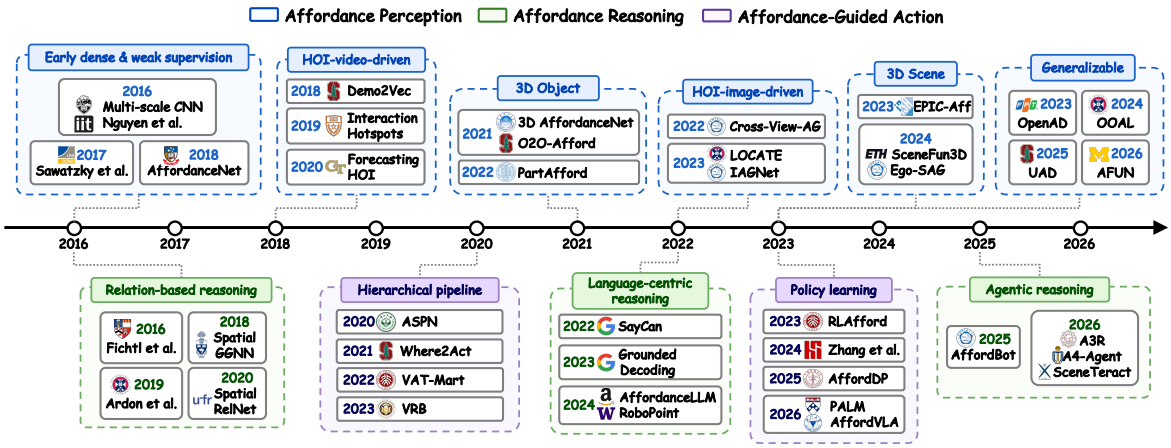


Figure 2. Temporal evolution of representative advances in affordance learning over the past decade (2016–2026).

From dense labels to diverse evidence. Early methods commonly relied on dense manually annotated supervision, such as pixel-level affordance masks over objects and their parts [2,4]. This setting provides clear training targets and evaluation, but it is costly to scale because annotators must reason about potential object uses and functional interactions rather than merely label visible boundaries. Alternative approaches draw on image-level labels, sparse points, human–object interactions, and model-generated pseudo-labels or priors [7,10,42,48,49]. Physical simulation provides further evidence by generating interaction labels for learning [50] or directly evaluating candidate interactions to infer affordances and functional poses [51,52]. These approaches reduce reliance on dense manual annotation, although their reliability depends on data quality, model priors, and simulation fidelity.

From 2D image plane to 3D metric space. Classical affordance prediction often represented action possibilities on the image plane, such as 2D masks, heatmaps, keypoints, or part-level regions. Although these predictions identify visible functional areas, they remain view-dependent and do not directly encode metric geometry. Embodied agents, however, act in 3D space and require estimates of surface position, orientation, scale, reachability, navigability, and surrounding constraints in the physical workspace. This need has driven affordance grounding from 2D image regions toward 3D physical space, as illustrated in Figure **??**(a). By grounding affordances directly in metric space, 3D-based

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

approaches associate functional regions with physical surfaces, integrate evidence across viewpoints, and support feasibility assessment and action planning.

From static regions to actionable dynamics. Early visual methods often represent affordances as binary or multi-class masks, indicating what action a region supports and where it applies. Later work additionally models how interaction should unfold: contact points identify candidate interaction locations [9,53], motion parameters describe part movement [6,54], poses specify how an agent's hand or end-effector should be configured for interaction [43,55], and post-contact trajectories describe behavior after contact [10,56]. The appropriate representation depends on its intended downstream use: masks and heatmaps are suitable for perceptual grounding, contact points and actionability scores support action selection, and motion directions, poses, or trajectories connect affordance with planning and control.

From closed-set recognition to open-set generalization. Classical supervised settings typically assume predefined objects, affordance labels, and action categories, with substantial overlap between training and test distributions. This closed-set assumption is restrictive in real environments containing long-tail objects, unseen object–action combinations, and instructions that do not match a fixed affordance vocabulary. As shown in Figure 3(b), the number of open-set affordance methods has increased substantially in recent years. Open-set methods address this limitation by using language, limited demonstrations, and foundation models to generalize beyond predefined labels. Open-vocabulary 3D approaches align language with point features to recognize unseen affordances and object categories [57–59], while vision-language grounding exploits instructions and multimodal priors to address novel objects, part queries, and interaction goals [13,46,60]. This direction reduces the need for exhaustive affordance annotation when embodied agents encounter new tasks and objects.

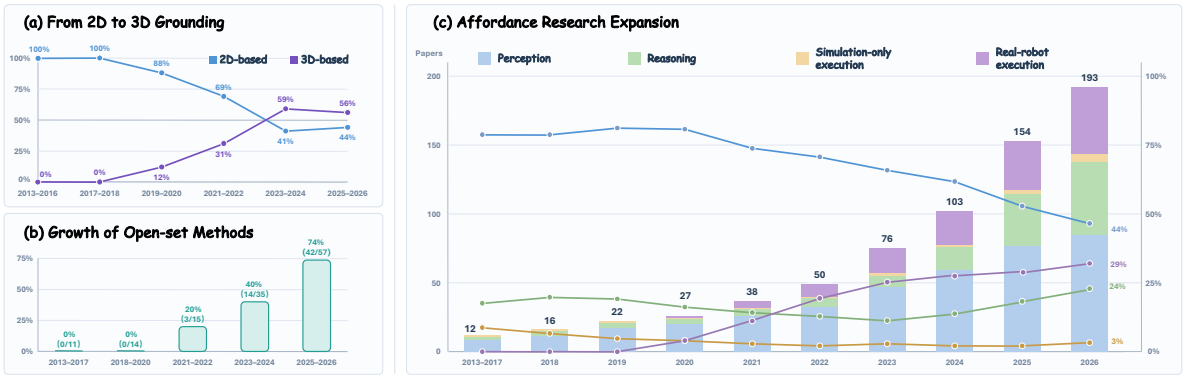


Figure 3. Research trends in affordance learning through July 2026. (a) and (b) analyze affordance grounding across perception and reasoning, showing the transition from 2D to 3D and the growth of open-set methods, respectively. (c) shows the cumulative publications and their distribution across perception, reasoning, simulation execution, and real-robot execution.

From modular pipelines to integrated policy learning. In robotics, affordance has often been used in a hierarchical pipeline that predicts affordance regions, plans a grasp or motion, and executes the resulting action. Such modularity is interpretable and can explicitly incorporate physical feasibility and robot kinematics, but hand-designed representations, planner interfaces, or primitives may limit transfer across tasks, objects, and embodiments. Recent systems increasingly integrate affordance into policy learning and action prediction: it may condition diffusion or vision-language-action policies explicitly [18,61,62], appear as an implicit intermediate representation [47,63], and provide auxiliary supervision or an optimization signal [64,65]. This integration aligns affordance learning more closely with the robot's embodiment, action space, and execution objective, shifting it from a standalone perceptual output toward a functional component of embodied decision making.

Together, these shifts show how affordance learning is becoming more scalable, task-conditioned, spatially grounded, and action-relevant. For embodied agents, these developments support identifying interaction opportunities, contextualizing them under task goals, and grounding them in feasible

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

actions. Figure 3(c) further shows rapid expansion of affordance research, with an increasing number of studies exploring embodied settings and real-world robot integration. Nevertheless, affordance grounding remains dominant: perception- and reasoning-oriented studies account for 68% of the 193 surveyed works, indicating that connecting affordance understanding with physical execution remains an open challenge.

3. Affordance Perception

Affordance perception addresses the grounding problem in affordance learning: given an observation, it identifies potential action opportunities and where their corresponding functional entities are located. We categorize a method as perception when its main contribution is to detect, segment, score, or spatially localize affordances through masks, heatmaps, keypoints, point-level scores, or 3D regions. Such grounding can be conditioned on action verbs, affordance concepts, or task queries, and can draw supervision or priors from observed human behavior or foundation models.

We organize affordance perception through four complementary and non-exclusive perspectives, as shown in Figure 4. The first two concern the grounding space. **Visual affordance perception** predicts functional regions on the 2D image plane, while **spatial affordance perception** grounds affordances in metric 3D representations of objects or scenes. **Interaction-driven affordance perception** derives supervision or contextual cues from human–object interaction images, videos, or task demonstrations, whereas **generalizable affordance perception** studies transfer to novel objects, unseen affordance labels, open-vocabulary queries, or unfamiliar interaction contexts. Table 1 summarizes representative methods across these perspectives, organized by setting, input modality, output form, supervision regime, and open-set generalization ability.

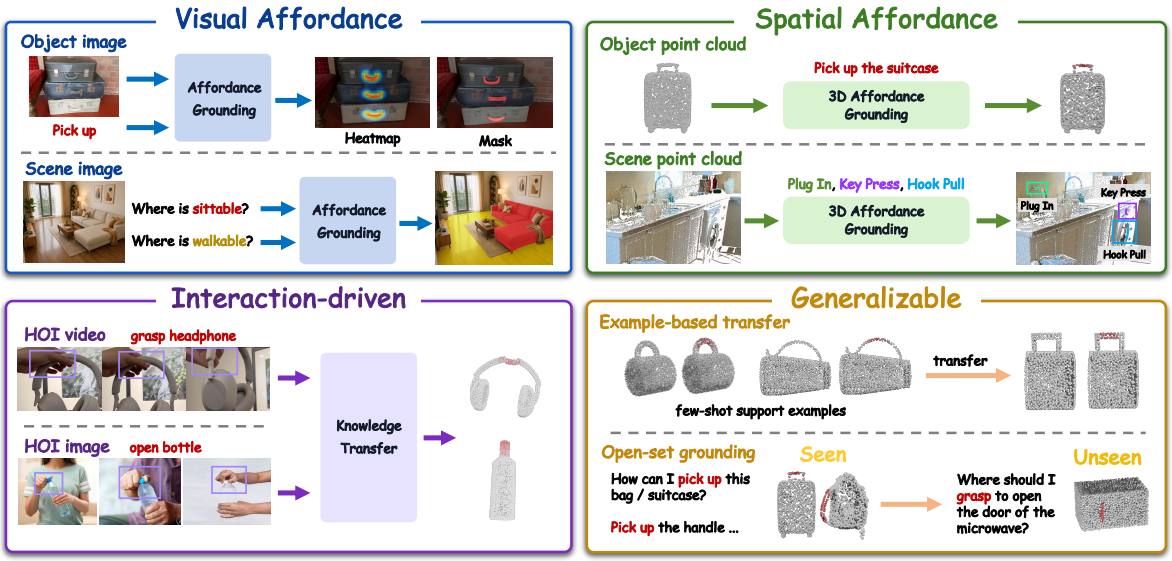


Figure 4. Taxonomy of affordance perception. Existing approaches are characterized through four complementary and non-exclusive perspectives. Visual and spatial affordance perception distinguish the grounding space; interaction-driven affordance perception focuses on supervision and contextual cues derived from observed behavior; and generalizable affordance perception addresses transfer beyond the training distribution.

3.1. Visual Affordance Perception

Visual affordance perception studies affordance grounding on the 2D image plane. Given an RGB or RGB-D observation, it localizes image regions associated with an interaction, where the interaction may be predicted from a predefined label set or specified by a task or query. Existing methods differ primarily in prediction scope: **object-centric grounding** identifies functional regions of individual objects, while **scene-level grounding** situates affordances within broader environmental layouts. Together, they provide fine-grained object-part cues and context-aware interaction opportunities for downstream action.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

Table 1. Comparison of representative affordance perception methods in Section 3.

Method	Venue & Year	Setting	Input Modality						Output				Supervision	Open-set
			I	D	V	3D	L	HOI	M	H	P	Mo		
Nguyen et al. [2]	IROS 2016	2D object	●	●	○	○	○	○	●	○	○	○	Strong	✗
Sawatzky et al. [48]	CVPR 2017	2D object	●	○	○	○	○	○	●	○	○	○	Weak	✗
AffordanceNet [4]	ICRA 2018	2D object	●	○	○	○	○	○	●	○	○	○	Strong	✗
Demo2Vec [7]	CVPR 2018	2D object	●	○	●	○	○	●	○	●	○	○	Strong	✗
Interaction Hotspots [42]	ICCV 2019	2D object	●	○	●	○	○	●	○	●	○	○	Weak	✗
Ego-Topo [66]	CVPR 2020	2D scene	○	○	●	○	○	●	○	○	○	○	Weak	✗
AffKp [67]	RA-L 2021	2D object	●	●	○	○	○	○	●	○	●	●	Strong	✗
OS-AD [68]	IJCAI 2021	2D object	●	○	○	○	○	●	●	○	○	○	Strong	✓
O2O-Afford [50]	CoRL 2021	3D object	○	○	○	●	○	○	○	●	○	○	Automatic	✗
Cross-view-AG [45]	CVPR 2022	2D object	●	○	○	○	○	●	○	●	○	○	Weak	✗
PartAfford [69]	ECCVW 2022	3D object	○	○	○	●	○	○	●	○	○	○	Weak	✗
AffCorrs [70]	CoRL 2022	2D object	●	○	○	○	○	○	●	○	○	○	None	✓
LOCATE [71]	CVPR 2023	2D object	●	○	○	○	○	●	○	●	○	○	Weak	✗
VRB [10]	CVPR 2023	2D object	●	○	●	○	○	●	○	●	○	●	Automatic	✗
IAGNet [72]	ICCV 2023	3D object	●	○	○	●	○	●	○	●	○	○	Strong	✗
OpenAD [57]	IROS 2023	3D object	○	○	○	●	●	○	●	○	○	○	Strong	✓
Where2Explore [73]	NeurIPS 2023	3D object	○	○	○	●	○	○	○	●	○	○	Automatic	✗
LEMON [74]	CVPR 2024	3D object	●	○	○	●	○	●	●	●	○	○	Strong	✗
PointRefer [75]	CVPR 2024	3D object	○	○	○	●	●	○	○	●	○	○	Strong	✓
OOAL [76]	CVPR 2024	2D object	●	○	○	○	●	○	●	●	○	○	Strong	✓
FAKP-Net [77]	IROS 2024	3D object	●	●	○	○	○	○	●	○	●	○	Strong	✗
INTRA [78]	ECCV 2024	2D object	●	○	○	○	●	●	○	●	○	○	Weak	✓
MaskPrompt [79]	AAAI 2025	2D object	●	○	○	○	●	○	●	○	○	○	Strong	✓
Reasoning Mamba [80]	CVPR 2025	2D object	●	○	○	○	○	●	○	●	○	○	Weak	✗
GEAL [81]	CVPR 2025	3D object	○	○	○	●	●	○	○	●	○	○	Strong	✓
GLANCE [82]	ICCV 2025	3D object	○	○	○	●	●	○	○	●	○	○	Strong	✓
LoopTrans [83]	ICCV 2025	2D object	●	○	○	○	○	●	○	●	○	○	Weak	✗
FunREC [84]	CVPR 2026	3D scene	○	●	●	○	○	●	●	○	○	●	None	✗
Affostruction [85]	CVPR 2026	3D object	●	●	○	○	●	○	○	●	○	○	Automatic	✓
AffordMatcher [86]	CVPR 2026	3D scene	●	○	○	●	○	●	●	○	○	○	Strong	✗
DAG [87]	ECCV 2026	3D object	●	○	○	●	●	●	○	●	○	○	Strong	✓
Affogato [88]	ECCV 2026	3D object	○	○	○	●	●	○	○	●	○	○	Automatic	✓
AffordanceSAM [89]	ACM MM 2026	2D object	●	○	○	○	●	○	○	●	○	○	Strong	✓

Input modalities are denoted by **I** (RGB image), **D** (depth), **V** (video), **3D** (e.g., point cloud), **L** (language/task query), and **HOI** (human–object interaction input). Output forms are denoted by **M** (discrete mask), **H** (continuous 2D/3D heatmap), **P** (keypoint), and **Mo** (motion, trajectory, or articulation parameters). Supervision regime is classified as **Strong** (dense spatial labels), **Weak** (sparse points/image-level labels), **Automatic** (interaction-, simulation-, or model-generated labels), or **None** (training-free methods built on frozen pretrained features). **Open-set** denotes the ability to recognize or ground both unseen object categories and unseen affordance concepts.

Early efforts mainly investigated object-centric affordance grounding from 2D visual observations. Nguyen et al. [2,3] develop convolutional approaches that first combine appearance with depth-derived cues and later decouple object localization from pixel-level affordance refinement. AffordanceNet [4] extends instance segmentation to jointly detect objects and predict fine-grained affordance masks in an end-to-end framework. Complementary work examines object-part structure and develops manipulation-oriented representations. Part-based analysis highlights the role of geometric object parts in tool functionality [1], while part–affordance coupling [90] and relationship-aware detection [91] model associations between parts and affordances and dependencies among affordances. Efficient attention-based segmentation [92] and boundary-preserving detection [93] further refine functional regions using contextual and boundary cues. Beyond region masks, AffKp [67] encodes manipulation positions and directions as structured 2D keypoints, while AffContext [94] uses predicted object affordances as intermediate cues for manipulation-oriented scene reasoning. To improve scalability, AffordanceLearn [95] reduces reliance on dense manual annotation through synthetic training data. Further efforts expand dataset scale and viewpoint coverage through a large-scale multi-view RGB-D affordance benchmark [96] and develop a hierarchical vision transformer to capture affordance cues across viewpoints [97].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

In parallel, scene-level perception broadens the prediction target from isolated objects to action-relevant regions in full visual contexts. Early approaches combine appearance, semantic structure, and spatial layout as complementary cues. Roy and Todorovic [98] infer depth, surface normals, and semantic scene structure from a single RGB image and combine these multi-scale cues for affordance segmentation. Luddecke and Worgotter [99] derive dense affordance supervision by mapping object- and part-level semantic annotations to affordance labels. A subsequent extension [100] expands this label-transfer formulation to a larger set of robot-relevant affordances and evaluates it in robotic settings. Qian and Fouhey [44] jointly predict object properties, action types, and affordance heatmaps from a single RGB image and query points. Panoramic methods further extend scene-level prediction to whole-environment observations. PanoAffordanceNet [101] studies holistic affordance prediction in 360-degree indoor scenes and introduces 360-AGD to capture affordance cues under equirectangular distortion, semantic dispersion, and cross-scale layout variation. PAP [102] adopts a coarse-to-fine pipeline to recover affordance masks from ultra-high-resolution panoramic imagery.

Orthogonal to the object–scene distinction, existing methods also differ in supervision strength. Most object-centric and scene-level approaches rely on dense pixel-level affordance annotations, which are costly because annotators must infer potential interactions rather than simply delineate visible object boundaries. **Weakly supervised affordance perception** reduces this burden by learning dense functional predictions from sparse spatial annotations or coarse interaction labels. One line replaces dense masks with sparse point supervision. Sawatzky et al. [48], Sawatzky and Gall [103] perform affordance segmentation from a limited number of annotated points, while STRAP [104] further propagates sparse supervision through label- and spatial-affinity constraints. A complementary line exploits action labels from human–object interaction data. Interaction Hotspots [42] learns from action-labeled human–object interaction videos, whereas Cross-view-AG [45] transfers affordance cues from exocentric interaction images to target objects. Both localize affordances from coarse action labels using CAM-style activation maps [105].

Overall, visual affordance perception grounds interaction opportunities in 2D images. Object-centric methods identify functional parts and structured interaction cues on individual objects, scene-level methods situate affordances within broader environmental contexts, and weakly supervised methods reduce reliance on dense pixel-level annotations. These settings establish image-plane affordance grounding as a fundamental component of embodied perception by identifying where interactions may be supported.

3.2. Spatial Affordance Perception

Spatial affordance perception grounds affordances in metric 3D representations, such as point clouds, meshes, and 3D Gaussian representations. Its key distinction from RGB-D image prediction lies in the output space. While depth provides geometric cues, spatial affordance perception directly associates affordance predictions with object surfaces or scene locations in 3D coordinates. This grounding makes contact locations, surface geometry, orientations, and spatial relations explicit, providing a geometric interface for downstream action. Existing methods span two main settings: **3D object grounding**, which localizes affordances on object geometries, and **3D scene grounding**, which anchors affordances to locations or functional elements in full 3D environments.

Early work on spatial affordance perception primarily addressed object-level grounding on point-cloud representations. Kokic et al. [106] predict part-level affordance scores from complete or partial object point clouds, jointly estimate object class and orientation, and convert these predictions into contact-location and approach-direction constraints for task-specific grasp planning. 3D AffordanceNet [5] later introduces the first large-scale 3D affordance benchmark with dense probabilistic point-wise annotations propagated from sparse affordance keypoints, supporting full-shape, partial-view, and rotation-invariant affordance estimation. Subsequent methods broaden 3D object grounding by reducing annotation requirements, improving geometric robustness, and producing more manipulation-oriented outputs. O2O-Afford [50] uses simulation to generate annotation-free supervision for object–object interaction regions, while PartAfford [69] discovers part-level affordances from

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

object-level labels through reconstruction and primitive-based regularization. LGAfford-Net [107] emphasizes local surface geometry, prototype-based detection [108] explains predictions through learned prototypes, and Affostruction [85] reconstructs complete geometry from partial RGB-D observations to ground affordances on both observed and unobserved surfaces. To make these predictions more directly actionable, ACF [109] estimates 3D keypoints and directed axes of functional parts; FAKP-Net [77] encodes execution position, direction, and extent as 3D keypoint quadruplets; and ComA [110] models human–object proximity and orientation distributions.

At the scene level, spatial affordance perception anchors interaction opportunities to persistent 3D environments rather than isolated object shapes. EPIC-Aff [111] aggregates egocentric interactions into multi-label affordance maps registered in scene coordinates, so that affordances become properties of metric locations and a single location may support several actions. SceneFun3D [6] provides fine-grained functional-element annotations, task descriptions, and motion parameters in real indoor scans, supporting functionality segmentation, task-driven affordance grounding, and 3D motion estimation. IAAO [112] constructs hierarchical feature and label fields on 3D Gaussian primitives and jointly estimates affordances, global transformations, and local articulation parameters. AffordMatcher [86] establishes semantic correspondences between image-based and point-cloud instances, using visual signifiers to identify affordance regions in 3D scenes.

Overall, spatial affordance perception moves affordance grounding from image coordinates to metric object and scene representations. Object-level methods localize functional surfaces and capture manipulation-relevant geometry, whereas scene-level methods associate interaction opportunities with locations and functional elements through 3D environments. Together, these 3D affordance estimates provide a spatial interface for downstream planning and embodied interaction.

3.3. Interaction-Driven Affordance Perception

Humans do not acquire affordance knowledge solely by memorizing object categories or explicitly labeled functional parts. Instead, much of this knowledge emerges from observing how people interact with objects. Through such observations, humans learn to associate specific object regions with particular actions and can readily transfer these expectations to novel instances. This learning mechanism motivates interaction-driven affordance perception, which derives supervision or contextual cues from observed human–object interactions (HOI) rather than relying solely on directly annotated affordance regions. Existing methods mainly pursue this idea using two forms of data: **demonstration or human–object interaction videos**, which reveal temporal dynamics such as hand motion, object state change, and action trajectories, and **human–object interaction images**, which provide scalable spatial cues about contact and functional regions.

Video-based methods draw on both exocentric and egocentric observations. Exocentric videos, collected from web demonstrations or third-person recordings, capture human poses, manipulated objects, and broader task context at scale [7,113]. This holistic view supports learning action-conditioned usage patterns, but can obscure fine-grained contact and introduces a viewpoint gap from embodied agents. Egocentric videos [114,115] better align with embodied perception and expose hand dynamics, object-state transitions, and upcoming contact, making them informative about where and how manipulation will occur. However, rapid camera motion, restricted fields of view, and embodiment-specific biases remain challenging. One line of work transfers temporal interaction evidence to static target images. Demo2Vec [7] encodes a product-demonstration video to predict the action and interaction region in a target image of the same object, thereby transferring action-conditioned visual priors from video to a static view. Afformer [116] improves this setting with a fine-grained decoder and MaskAHand pre-training to bridge the video–image discrepancy. These methods transfer action-conditioned cues to static images but still require annotated heatmaps and action labels. Interaction Hotspots [42] instead uses action-labeled HOI videos as weak supervision, transferring features from active interactions to infer action-specific hotspots on inactive object images. Other work mines structured cues from egocentric video: Liu et al. [8] automatically extract hand trajectories and future hotspots and train an object-centric transformer to anticipate interaction; VRB [10] learns contact points and post-contact

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

trajectories as a reusable representation for robot learning; Yoshida et al. [117] add textual annotations and rotational motion; and Li et al. [118] derive graspable and functional points as segmentation-mask supervision for broader data collection. More recently, Chen et al. [119] learn action, dynamics, and value models from human videos, capturing task-conditioned predictions of how actions transform the scene and contribute to task completion. For a broader review of human-video supervision for robotics, we refer to Feng et al. [120].

In parallel, a larger body of work learns from human–object interaction images, which provide less temporal information but are easier to collect at scale. Cross-view-AG [45] first formulates exocentric-to-egocentric affordance grounding, using interaction images as low-cost supervision without pixel-level affordance masks. This makes HOI images practical affordance supervision, although human occlusion and action ambiguity complicate the transfer. Subsequent methods extract more reliable object-centric and part-level signals: Luo et al. [121] model relations between human body parts and object regions, whereas LOCATE [71] clusters interaction-region features into prototypes and selects the object-part prototype for target grounding. Later methods enrich this evidence with textual action descriptions [122], depth and semantic prompts [123], or part-level semantic priors [124]. Other approaches improve representation learning and transfer. INTRA [78] learns interaction-aware representations without paired ego–exo images; Reasoning Mamba [80] models local and global visual relations through hypergraph-guided state-space learning; LoopTrans [83] uses a closed loop to refine exocentric knowledge; and Moon et al. [125] introduce prototypical and pixel-level contrastive learning to capture affordance-relevant cues at both the object and part levels.

Interaction evidence can also condition affordance grounding in 3D, transferring human-centered visual cues to metric object representations. IAGNet [72] aligns a 2D interaction image with point-cloud object regions for point-level prediction, while LEMON [74] jointly predicts object affordances, human contact, and their 3D spatial relations from an interaction image and paired geometries. MIFAG [126] fuses invariant patterns from multiple interaction images with point-cloud features, and VideoAfford [127] extracts temporal priors from HOI videos using a multimodal large language model. Beyond individual objects, observed interactions reveal how scene regions are used. Ego-Topo [66] organizes egocentric video into activity-centric zones, while EPIC-Aff [111] projects first-person interactions into metric maps where multiple affordances may coexist. Ego-SAG [128] aligns egocentric interaction video with a paired 3D scene point cloud to predict point-level affordance masks. FunREC [84] goes beyond grounding by reconstructing simulation-ready functional scenes and recovering articulated geometry, motion, and kinematics from RGB-D video, while AffordMatcher [86] uses HOI images as visual signifiers for locating functional object parts in 3D scenes.

Overall, interaction-driven affordance perception converts observed behavior into supervision and contextual priors for affordance grounding. Videos provide temporal contact and motion cues, static HOI images offer scalable cross-view evidence, and cross-modal methods transfer these cues to 3D objects and environments. Together, these approaches make human behavior a practical source of affordance supervision across representations.

3.4. Generalizable Affordance Perception

Most affordance-perception methods assume predefined affordance categories, object distributions, and task settings. Embodied agents, however, must handle objects absent from training, unseen action expressions, and unfamiliar interaction contexts. Generalizable affordance perception addresses these cases without requiring dense annotations for every object–affordance combination. Existing work mainly follows two complementary paradigms: **example-based transfer**, which uses sparse support observations, annotated samples, or reference interactions as task-specific evidence; and **open-set grounding**, which uses language and foundation-model knowledge to move beyond fixed label spaces.

Example-based transfer covers support-conditioned grounding and learning or adaptation from one or a few examples. In support-conditioned grounding, a reference observation specifies the affordance to be transferred to a target. One-Shot Affordance Detection [68,129] uses an HOI support

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

image to indicate an action purpose and segments complete target objects sharing that affordance. AffCorrs [70] instead performs training-free part-level transfer by matching an annotated reference region with corresponding parts in a target scene through cyclic correspondence. Other methods use limited labeled examples as sparse training supervision. OOAL [76] is a representative approach that learns from one labeled image per base object category and exploits foundation-model priors for zero-shot grounding of novel object and affordance classes. OS-AGDO [130] extends one-shot grounding to deformable objects by combining semantic and geometric cues to accommodate changes in shape, appearance, and configuration. Example-based transfer also applies to 3D and interactive settings. Where2Explore [73] adapts to unseen articulated-object categories through a few exploratory interactions, transferring known affordances via local geometric similarity while selecting informative regions to explore. O3Afford [131] learns bidirectional object-to-object affordance maps from one annotated object pair per affordance type by combining pretrained visual semantics with point-cloud geometry. AffoRo-GS [132] grounds verb-conditioned affordances in 3D Gaussian scene representations from multi-view observations using only one affordance-labeled image per scene.

Open-set grounding uses semantic queries and pretrained knowledge to generalize beyond fixed vocabularies and training distributions. In 2D, MaskPrompt [79] refines object-shape masks to enhance visual features and construct mask-focused textual prompts for open-vocabulary affordance segmentation. AffordanceSAM [89] adapts a text-promptable SAM backbone through an affordance module and coarse-to-fine supervised training, supporting transfer to novel objects and action queries. UAD [49] uses large vision and vision-language models to generate instruction–affordance pairs automatically from rendered objects, then distills them into a lightweight task-conditioned predictor without manual affordance annotations. In 3D, several methods directly align language with point-level geometry. OpenAD [57] jointly embeds affordance text and point features for zero-shot detection of unseen labels, while Van Vo et al. [59] improve transfer to unseen labels and objects through knowledge distillation and explicit text–point correlation. 3DAPNet [58] further predicts compatible 6-DoF poses together with affordance regions. PointRefer [75] grounds object regions specified by natural-language functional queries, and GLANCE [82] adds intermediate cross-modal connectors and geometric priors to improve generalization to unseen object categories. OVA-Fields [133] extends this paradigm to sparse and noisy RGB-D scene reconstructions, combining multilevel geometry with semantic affordance embeddings to map complex commands to operational parts under weak supervision. Foundation models further improve the scale and robustness of 3D grounding. GEAL [81] transfers pretrained 2D semantic priors to point clouds through Gaussian-splatting-based cross-modal consistency, improving generalization to novel categories and corrupted observations. Affogato [88] orchestrates vision-language and segmentation models to generate open-vocabulary queries and corresponding 3D affordance heatmaps at scale. DAG [87] extracts multi-level affordance priors from a frozen text-to-image diffusion model, combines them with affordance text, and fuses the result with point-cloud features for dense 3D prediction.

Overall, example-based methods specify target functions through sparse visual or interaction evidence, whereas open-set methods use language and pretrained priors to broaden the objects and affordances that can be grounded. Recent LLM- and MLLM-based approaches further draw on semantic and commonsense knowledge to acquire generalization capabilities, but often require instruction understanding, intent inference, and task-relevant affordance selection. We therefore discuss these reasoning aspects in Secs. 4.2 and 4.3. Together, these directions extend affordance perception to long-tail objects, novel queries, and interaction contexts that cannot be exhaustively annotated in advance.

4. Affordance Reasoning

Affordance reasoning studies how an embodied agent determines which affordances are appropriate based on task requirements, object relations, scene context, agent state, or available skills. It resolves ambiguity among possible interactions through relational structure, contextual evidence, or

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

knowledge from large pretrained models. This distinguishes it from affordance perception, which directly grounds candidate interaction regions from observations. For language-conditioned methods, a query used only to localize a region remains perception, whereas a method that leverages knowledge from LLMs or MLLMs, decomposes instructions, or interprets implicit intent to select the relevant affordance is treated as reasoning.

We organize affordance reasoning into three complementary paradigms according to their primary reasoning mechanism, as summarized in Figure 5. **Relation-based affordance reasoning** models functional knowledge explicitly through semantic or structural relations among objects, parts, actions, attributes, or scene context. **Language-centric affordance reasoning** uses LLMs or MLLMs to draw on model knowledge, decompose instructions, or interpret intent when determining the relevant affordance. **Agentic affordance reasoning** further frames the reasoning process as a structured workflow that coordinates specialized tools, evidence acquisition, and intermediate validation or refinement. Table 2 summarizes representative methods across these paradigms.

Table 2. Comparison of representative affordance reasoning methods in Section 4.

Paradigm	Method	Venue & Year	Setting	Text Form	Output Form
Relation-based	Zhu et al. [134]	ECCV 2014	2D object	–	Affordance label
	Fichtl et al. [135]	TCDS 2016	3D object	–	Relational affordance
	Chuang et al. [136]	CVPR 2018	2D scene	–	Affordance label, explanation
	Ardón et al. [137]	RA-L 2019	2D scene	–	Grasp affordance
	Spatial-RelNet [138]	ICRA 2020	2D scene	–	Placement affordance
	AR-Net [139]	arXiv 2021	2D scene	–	Relational affordance
	Xin et al. [140]	ICDH 2022	2D scene	–	Affordance label
Language-centric	SayCan [11]	CoRL 2022	2D scene	Implicit intent	Action sequence
	Grounded Decoding [12]	NeurIPS 2023	2D scene	Implicit intent	Action sequence
	AffordanceLLM [13]	CVPRW 2024	2D object	Action term	Heatmap
	RoboPoint [60]	CoRL 2024	2D scene	Explicit instruction	Keypoints
	ManipVQA [141]	IROS 2024	2D scene	Explicit instruction	Bounding box, mask
	WorldAfford [46]	ICTAI 2024	2D scene	Implicit intent	Heatmap
	OVAL-Prompt [142]	ICRAW 2024	2D object	Action term	Mask
	3D-ADLLM [143]	ICLR 2025	3D object	Explicit instruction	Mask
	Afford-X [144]	arXiv 2025	2D object	Action term	Bounding box, mask
	SeqAfford [145]	CVPR 2025	3D object	Implicit intent	Heatmap sequence
	LMAffordance3D [146]	CVPR 2025	3D object	Explicit instruction	Heatmap
	GREAT [147]	CVPR 2025	3D object	Explicit instruction	Heatmap
	Fun3DU [148]	CVPR 2025	3D scene	Explicit instruction	Mask
	ViSPLA [149]	NeurIPS 2025	3D object	Explicit instruction	Mask
	PAVLM [150]	IROS 2025	3D object	Explicit instruction	Heatmap
	Aff3DFunc [151]	ACM MM 2025	3D object	Explicit instruction	Heatmap
	SeqAffordSplat [152]	arXiv 2025	3D scene	Implicit intent	Mask sequence
	AffordGrasp [153]	IROS 2025	2D scene	Implicit intent	Mask
	3DAffordSplat [154]	ACM MM 2025	3D object	Explicit instruction	Mask
	Affordance-R1 [155]	AAAI 2026	2D object	Implicit intent	Mask
	TASA [156]	AAAI 2026	3D scene	Explicit instruction	Mask
	TokAG [157]	ECCV 2026	2D object	Explicit instruction	Heatmap
Agentic	ASP [158]	arXiv 2025	3D scene	Explicit instruction	Mask, action sequence
	AffordBot [54]	NeurIPS 2025	3D scene	Explicit instruction	Mask, motion type/axis
	SceneTeract [159]	arXiv 2026	3D scene	Implicit intent	Feasibility diagnostics
	A3R [160]	arXiv 2026	3D scene	Explicit instruction	Mask
	A-Harness [161]	arXiv 2026	2D scene	Implicit intent	Mask
	AffordMEM [162]	arXiv 2026	3D scene	Explicit instruction	Mask
	A4-Agent [163]	ECCV 2026	2D object	Implicit intent	Mask
	PAP [102]	ECCV 2026	2D scene	Implicit intent	Mask

Text form is classified as **Action term** (a verb or short action phrase), **Explicit instruction** (an instruction or query that explicitly specifies the target object or interaction), or **Implicit intent** (a goal that leaves the target object and interaction region unstated); – indicates no direct language input. Output form lists the principal affordance predictions exposed by each method.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

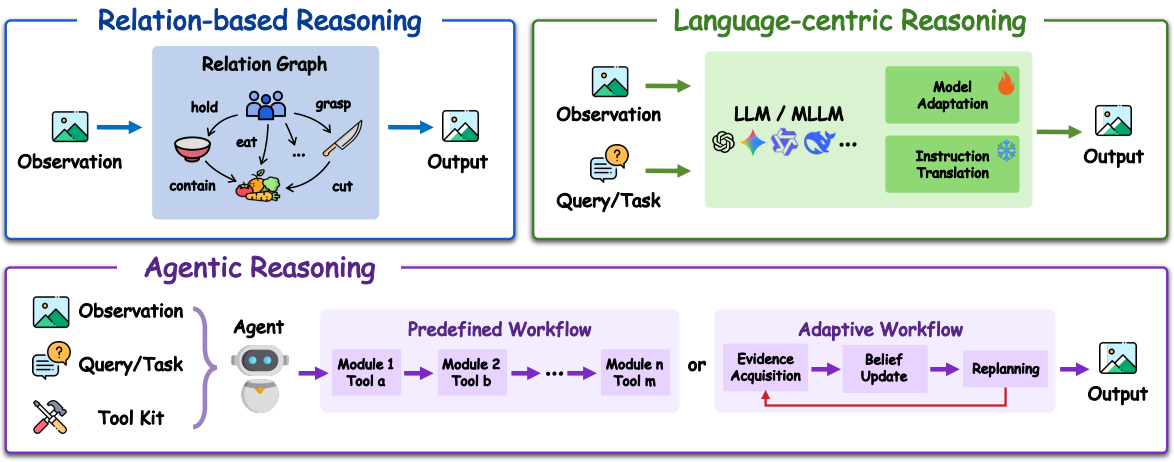


Figure 5. Taxonomy of affordance reasoning. Existing approaches are organized into three paradigms: relation-based affordance reasoning, language-centric affordance reasoning, and agentic affordance reasoning.

4.1. Relation-Based Affordance Reasoning

Relation-based affordance reasoning represents an early route for resolving functional ambiguity before the LLM era. Its central idea is to treat affordance as an explicit relation over a predefined set of objects, parts, attributes, actions, or scene configurations. Rather than inferring affordances from isolated object appearance, these methods use semantic properties, object pairings, and surrounding context. Existing approaches encode such dependencies through probabilistic knowledge bases, learned spatial relations, and structured scene graphs.

One line encodes affordance knowledge through probabilistic logic. Zhu et al. [134] harvest object, attribute, and affordance facts from images and metadata and encode them in a knowledge base modeled with a Markov logic network. The resulting representation supports multiple visual inference tasks within a unified framework, including zero-shot affordance prediction. Ardón et al. [137] apply a related formulation to task-oriented grasping, using semantic attributes and context to infer a probability distribution over multiple grasp affordances. The inferred affordances are then associated with prototypical grasping patches on the object, connecting semantic reasoning with spatial grasp selection.

Other approaches focus on learning functional and spatial relations. Fichtl et al. [135] learn relational affordances from robot interaction and bootstrap new object-pair predictors using previously acquired affordance knowledge, reducing the experience required to learn new relations. Related work models spatial relations themselves as constraints on feasible interactions. Mees et al. [164] learn a metric over pairwise spatial relations that generalizes to novel object shapes and sizes from few examples. Spatial-RelNet [138] predicts pixel-wise placement probabilities conditioned on spatial relations, while Mees and Burgard [165] extend this model with language grounding and ambiguity resolution to compose pick-and-place commands into tabletop manipulation tasks. Graph-based models organize scene entities and their dependencies into explicit structures for contextual affordance inference. Chuang et al. [136] adopt graph neural networks to predict and explain whether an action–object pairing is physically appropriate or socially acceptable in the current situation. VAR-Net [140] combines semantic and visual relations among scene objects through graph attention. AR-Net [139] constructs a manipulation relationship graph from raw visual observations, using object attributes and subgraph context to predict manipulation relations between objects.

Taken together, relation-based methods show that appropriate affordances depend on object relations and scene context rather than isolated object recognition. Their explicit relational structure supports interpretable inference, but predefined entities and relation schemas can limit their flexibility under open-ended instructions and implicit goals.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

4.2. Language-Centric Affordance Reasoning

Language-centric affordance reasoning uses language, typically through LLMs or MLLMs, to decide the affordance target most relevant to a task. Unlike language-conditioned affordance perception, where text primarily specifies what region to localize, these methods use model knowledge, instruction decomposition, or intent interpretation to determine which object, part, function, or interaction constraint should be grounded. Early systems grounded language reasoning at the skill level: SayCan [11] combines the linguistic plausibility of candidate skills with their estimated executability, while Grounded Decoding [12] injects observation- and state-grounded scores directly into action generation. Subsequent work extends this principle from discrete skill selection to spatial affordance prediction.

Recent methods follow two main paradigms. The first integrates LLM- or MLLM-based reasoning directly into affordance prediction, adapting model knowledge through instruction tuning, multimodal alignment, distillation, or post-training. In 2D, AffordanceLLM [13] and RoboPoint [60] transfer vision–language knowledge to action-conditioned heatmaps or keypoints. Afford-X [144] explores a compact end-to-end model, and Affordance-R1 and AffordanceGrasp-R1 [155,166] use reward-based post-training to strengthen task-oriented reasoning. These approaches differ in output and adaptation strategy, but share the goal of embedding task interpretation into the spatial predictor rather than treating language as a fixed class label. This direction is supported by instruction-driven segmentation, grasping, and visual-question-answering datasets [141,167–171], which are discussed further in Section 6. Integrated reasoning has also expanded toward geometry-aware prediction. For 3D grounding, 3D-ADLLM [143] couples an LLM with a point-level decoder for free-form reasoning queries, ViSPLA [149] iteratively refines such predictions using geometric feedback, and PAVLM [150] aligns point-cloud features with contextual language embeddings. GREAT [147] further uses affordance chain-of-thought supervision to infer geometric properties and interaction intent from interaction images.

The second paradigm separates semantic reasoning from spatial grounding. An LLM or MLLM first translates instructions into task semantics, functional queries, part hypotheses, or augmented textual representations, while delegating spatial grounding to specialized modules. WorldAfford [46] derives affordance knowledge from free-form instructions, while OVAL-Prompt [142] explicitly queries the functional part before segmentation. AffordGrasp [153] further resolves implicit user goals into a target object and part for grasp generation. In 3D, part-aware grounding [172] and Aff3DFunc [151] align enriched functional text with part-level geometry; Fun3DU [148] reasons about the implied object across views and lifts its masks into a scene; and TASA [156] validates predicted concepts through reverse and fallback queries before segmentation. This decomposition preserves specialized spatial representations and makes intermediate semantics more explicit, although errors in the language-derived hypothesis may propagate to the grounder.

A further extension addresses sequential tasks, in which each action changes the scene and hence the next relevant affordance. NaturalVLM [173] selects a fine-grained step from the current observation, high-level instruction, and end-effector state before predicting its affordance. SeqAfford and SeqAffordSplat [145,152] predict sequences of 3D affordance regions in point-cloud and Gaussian-splatting representations, respectively, while TRACER [174] hierarchically decomposes intentions and refines functional regions on deformable objects. Unlike single-step grounding, these sequential methods make affordance selection conditional on task progress and changing state, providing a bridge toward long-horizon interaction.

Overall, integrated methods jointly learn task interpretation and spatial prediction, modular methods translate intent into queries or hypotheses for specialized grounders, and sequential methods update affordance decisions as tasks unfold. Language-centric reasoning thus provides a flexible semantic interface between task intent and spatial affordance grounding.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

4.3. Agentic Affordance Reasoning

Agentic affordance reasoning organizes affordance inference as an explicit workflow rather than a single model prediction. It decomposes the reasoning process into coordinated operations and invokes specialized tools or modules for functions such as visual grounding, scene querying, memory retrieval, and geometric verification. Compared with language-centric reasoning, which primarily relies on model knowledge to decide which affordance area should be grounded, agentic reasoning also exposes how this judgment is constructed and how intermediate evidence is acquired. Existing methods can be broadly organized into two regimes: **predefined workflows** connect specialized operations in a largely fixed sequence, whereas **adaptive workflows** condition subsequent tool calls, retries, replanning, or termination on accumulated evidence, belief states, or execution feedback. The distinction lies in runtime control flow rather than the number of modules.

Predefined workflows apply a fixed sequence of specialized operations. A4-Agent [163] separates zero-shot prediction into a Dreamer that generates a plausible interaction image, a Thinker that identifies the functionally relevant object part, and a Spotter that converts the resulting description into boxes, keypoints, and a precise affordance mask. AffordBot [54] segments 3D affordance-element candidates and projects their identifiers and types into pre-rendered surround views. An MLLM then selects an instruction-relevant view, identifies the referenced candidates, and estimates their motion type and discretized motion axis. PAP [102] adapts staged reasoning to ultra-high-resolution panoramas, using recursive visual routing to narrow the search space, adaptive gaze to correct local distortion, and cascaded grounding to recover the affordance mask. Some predefined workflows add memory or feasibility checks while retaining fixed control flow. AffordMEM [162] uses cross-scene memory to retrieve examples of small functional regions and in-scene spatial memory to organize candidate instances and their relations for resolving spatial references. SceneTeract [159] instead grounds task decomposition in an agent profile and maps each atomic action to relevant navigability, reachability, interactability, clearance, and visibility checks. It thereby turns a plausible interaction into an agent-conditioned feasibility judgment with step-wise diagnostics. These methods make intermediate decisions more transparent and task-specific, although their reasoning trajectories remain largely determined by manually designed module compositions and execution orders.

Adaptive workflows allow intermediate outcomes to determine what the system does next. Agentic Scene Policies [158] expose open-vocabulary retrieval, spatial-relation queries, and affordance-mediated interactions as tools over a queryable object map. Tool outputs and explicit failure messages are returned to the LLM to guide subsequent tool calls and replanning, connecting explicit affordance reasoning with policy execution. A3R [160] places the adaptive loop within evidence acquisition. It maintains a scene-grounded state comprising the current 3D Gaussian workspace, an affordance belief map, and the action-evidence history. Conditioned on this state, an MLLM iteratively selects 3D instance or part analysis, 2D semantic evidence with 3D lifting, workspace zoom, or termination. Each result updates the corresponding workspace, belief, or history state before the next decision, and the process ends when the policy selects termination or reaches its evidence-acquisition budget. A-Harness [161] instead instantiates a verification-driven adaptive loop. Its evidence store records heterogeneous tool outputs and their provenance, memory provides episodic priors, and a router selects and parameterizes the next skill. An affordance-specific verifier then checks cross-skill consistency, cross-scale stability, and evidence sufficiency. Failed verification identifies the missing or conflicting evidence and triggers a targeted corrective call, while successful verification allows the accumulated evidence to be fused into the final affordance prediction.

Overall, predefined and adaptive workflows form a continuum of agentic affordance reasoning. Predefined methods expose how an affordance judgment is assembled, while adaptive methods additionally revise the inference process using accumulated evidence or feedback. Agentic affordance reasoning therefore concerns how an affordance hypothesis is constructed, supported, and validated before commitment. Once the committed affordance is converted into embodiment-specific motion or closed-loop control, the focus shifts to affordance-guided action.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

5. Affordance-Guided Action

Affordance-guided action studies how grounded or inferred affordances are translated into executable robot behavior. At this stage, affordance serves as an interface between task-relevant interaction information and embodiment-specific action generation, providing cues such as interaction types, contact regions, poses, motion directions, trajectories, or spatial value fields. Affordance perception grounds candidate interactions, affordance reasoning determines which affordance is relevant to the task or context, and affordance-guided action uses the resulting information to generate or learn actions under robot-specific execution constraints.

We organize existing methods into two complementary paradigms according to how affordance enters action generation, as shown in Figure 6. **Hierarchical affordance-to-action** methods preserve affordance as an intermediate interface and pass it to a downstream grasp generator, controller, or planner to generate executable robot actions. **Affordance-integrated policy learning** instead incorporates affordance into a learned policy or action model as an explicit input, implicit representation, or optimization signal. These paradigms capture the primary way in which affordance contributes to action generation rather than mutually exclusive system architectures. Table 3 summarizes representative methods and the specific role affordance plays in each.

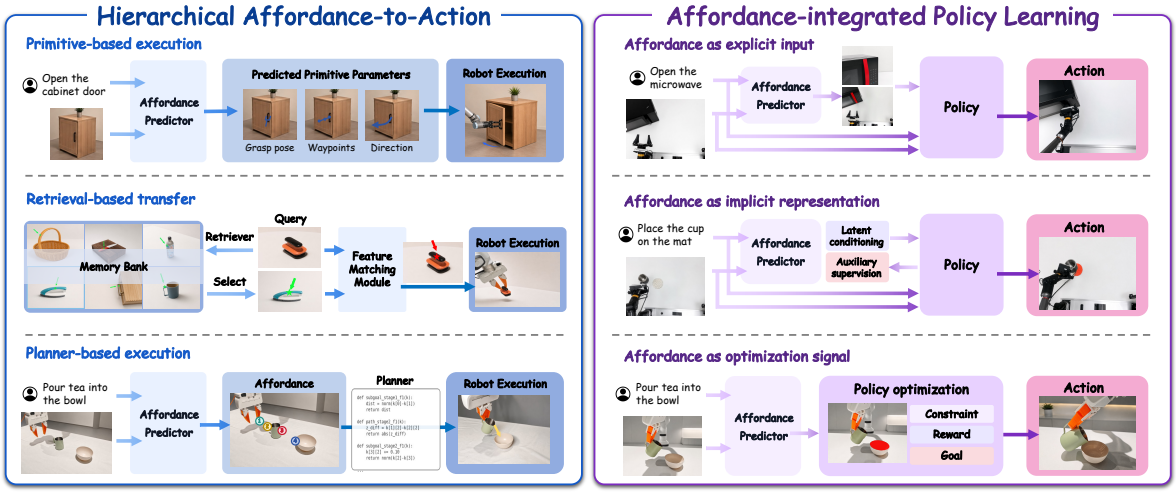


Figure 6. Taxonomy of affordance-guided action. Existing approaches are organized into two paradigms: hierarchical affordance-to-action, comprising primitive-based affordance execution, retrieval-based affordance transfer, and planner-based execution; and affordance-integrated policy learning, where affordance serves as an explicit input, implicit representation, or optimization signal.

5.1. Hierarchical Affordance-to-Action

Hierarchical affordance-to-action studies use affordance as an explicit interface between perceptual grounding and robot execution. Rather than mapping observations directly to motor commands, they first derive a task-relevant functional structure and then use it to guide downstream grasp generation, motion planning, trajectory optimization, or closed-loop control. Existing methods instantiate this idea through three complementary strategies: **primitive-based execution** uses predicted affordances to select grasps or parameterize motion primitives; **retrieval-based transfer** retrieves interaction knowledge from prior examples and adapts it to new objects or scenes; and **planner-based execution** treats affordance regions, fields, keypoints, or relations as targets or constraints for a downstream planner or optimizer. This modularity preserves interpretability while allowing robot-specific requirements, including kinematics, reachability, and collision avoidance, to be handled during execution.

Following the general formulation in Equation (1), hierarchical affordance-to-action can be expressed as an explicit two-stage mapping from affordance grounding to robot execution:

$$\mathcal{F} = f_\theta(o, l, c) = (a, r, \eta), \quad u = g_{\text{exec}}(\mathcal{F}, \mathcal{C}_{\text{robot}}). \tag{2}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

Here, f_θ derives the explicit affordance $\mathcal{F}$ from observation o, task information l, and optional context c, such as demonstrations, retrieved memories, or correspondence cues. Within $\mathcal{F}$, a, r, and η denote the interaction type, grounded region, and execution-related information (e.g., an action score, motion direction, or value field), respectively. The execution module g_{exec} then maps $\mathcal{F}$ to an action u under robot constraints $\mathcal{C}_{\text{robot}}$, and may take the form of a grasp generator, planner, controller, or optimizer that operates separately from the affordance predictor.

Table 3. Comparison of representative affordance-guided action methods in Section 5.

Paradigm	Method	Venue & Year	Input Modality				Affordance Role	Robot Setting
			I	D	3D	L		
Hierarchical	ASPN [175]	ICRA 2020	●	●	○	○	Primitive	Gripper
	Where2Act [176]	ICCV 2021	●	●	●	○	Primitive	Gripper
	VAT-Mart [177]	ICLR 2022	○	○	●	○	Primitive	Gripper
	AdaAfford [178]	ECCV 2022	○	○	●	○	Primitive	Gripper
	DualAfford [179]	ICLR 2023	○	○	●	○	Primitive	Dual Arm
	Wu et al. [180]	NeurIPS 2023	○	○	●	○	Primitive	Gripper
	DefoAfford [181]	ICCV 2023	●	○	○	○	Primitive	Gripper
	HULC++ [15]	ICRA 2023	●	○	○	●	Planning	Gripper
	VRB [10]	CVPR 2023	●	○	○	○	Primitive	Gripper
	LERF-TOGO [55]	CoRL 2023	●	●	○	●	Primitive	Gripper
	ACE-NBV [182]	CoRL 2023	●	●	○	○	Planning	Gripper
	VoxPoser [16]	CoRL 2023	●	●	○	●	Planning	Gripper
	DINOBot [183]	ICRA 2024	●	●	○	●	Retrieval	Gripper
	Robo-ABC [9]	ECCV 2024	●	●	○	○	Retrieval	Gripper
	RAM [53]	CoRL 2024	●	●	○	●	Retrieval	Gripper
	PreAfford [184]	IROS 2024	●	●	○	○	Planning	Gripper
	ReKep [17]	CoRL 2024	●	●	○	●	Planning	Gripper
	A_0 [185]	ICCV 2025	●	●	○	●	Primitive	Gripper
	Aff-Grasp [118]	ICCV 2025	●	●	○	●	Primitive	Gripper
	ManipGPT [186]	IROS 2025	●	○	○	●	Primitive	Gripper
	AffordDexGrasp [187]	ICCV 2025	●	○	●	●	Primitive	Dex. Hand
	ScaleADFG [188]	RA-L 2025	○	○	●	○	Primitive	Dex. Hand
	BiAssemble [189]	ICML 2025	○	○	●	○	Primitive	Dual Arm
	GarmentPile [190]	CVPR 2025	○	○	●	○	Primitive	Gripper
	O3Afford [131]	CoRL 2025	○	○	●	●	Planning	Gripper
	A3D [191]	AAAI 2026	○	○	●	○	Primitive	Dual Arm
	BiPreManip [192]	CVPR 2026	●	●	○	●	Primitive	Dual Arm
	RAAP [193]	ICRA 2026	●	●	○	●	Retrieval	Gripper
	FSAG [194]	IROS 2026	●	●	○	●	Planning	Dex. Hand
	AFI [195]	CVPR 2026	●	●	○	●	Planning	Gripper
Policy Learning	VAL [196]	ICRA 2021	●	○	○	○	Goal	Gripper
	Mandikal et al. [197]	ICRA 2021	●	●	○	○	Reward	Dex. Hand
	Borja-Diaz et al. [14]	ICRA 2022	●	●	○	○	Reward	Gripper
	RLAfford [198]	ICRA 2023	○	○	●	○	Reward	Gripper
	Bharadhwaj et al. [199]	ICRA 2023	●	○	○	○	Goal	Gripper
	AffordDP [61]	CVPR 2025	●	●	○	●	Explicit	Gripper
	DART [200]	T-Mech 2025	○	○	●	●	Explicit	Gripper
	CoA-VLA [62]	ICCV 2025	●	○	○	●	Explicit	Gripper
	RT-Affordance [18]	ICRA 2025	●	○	○	●	Explicit	Gripper
	UAD [49]	ICRA 2025	●	○	○	●	Explicit	Gripper
	GLOVER++ [201]	CoRL 2025	●	●	○	●	Explicit	Gripper
	DORA [202]	CBS 2025	○	○	●	○	Reward	Dex. Hand
	AffordDex [203]	AAAI 2026	○	○	●	○	Reward	Dex. Hand
	PALM [63]	CVPR 2026	●	○	○	●	Implicit	Gripper
	Affordance2Action [204]	arXiv 2026	●	○	○	●	Explicit	Gripper
	AffordVLA [47]	arXiv 2026	●	○	○	●	Implicit	Gripper
	Afford-VLA [64]	arXiv 2026	●	○	○	●	Implicit	Gripper
	AffordanceVLA [65]	arXiv 2026	●	○	●	●	Implicit	Gripper

Input modalities are denoted by **I** (RGB image), **D** (depth), **3D** (e.g., point cloud), and **L** (language/task query). **Affordance Role** indicates the primary role of affordance in each method.

Primitive-based execution converts affordance predictions into actions through grasp estimators or predefined motion primitives. ASPN [175] predicts a 3D affordance map for push and grasp selection without explicit hand–eye calibration, whereas Where2Act [176] couples point-level actionability with predefined action primitives for articulated objects. For task-oriented grasping, Chen et al. [205] combine coarse 6-DoF candidates with an affordance heatmap and grasp confidence to favor functionally relevant regions. AdaAfford [178] adapts predictions through a few test-time interactions

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

that reveal hidden kinematic and dynamic properties, while environment-aware affordance learning [180] incorporates occlusion, surrounding geometry, and robot morphology into actionability estimation. Subsequent methods enrich this interface with temporal and task-dependent action information. VAT-Mart [177] predicts contact locations together with action trajectories for articulated-object manipulation, shifting the representation from where to touch toward how the interaction should unfold after contact, while DefoAfford [181] makes a similar move for deformable objects by learning dense foresightful affordances that anticipate future object states. VRB [10] learns contact points and post-contact trajectories from human videos, projects these cues into 3D, and uses them to parameterize robot actions for imitation learning, exploration, goal-conditioned learning, and reinforcement learning. A_0 [185] similarly abstracts contact and post-contact motion into object-centric waypoints for cross-embodiment execution. PRIMP [206] adapts probabilistic motion primitives from simulated functional poses for obstacle-aware execution, whereas BridgeACT [207] couples tool–target grasp regions with 3D interaction dynamics for closed-loop control. Language and foundation models broaden primitive-based execution to novel objects and task instructions. LERF-TOGO [55] ranks grasp candidates using language-conditioned 3D relevance, whereas open-vocabulary part-based grasping [208] grounds task-relevant parts for 6-DoF grasp prediction. OVAL-Grasp [209] further filters candidates through grasp and avoidance regions, while ManipGPT [186] translates prompted affordance masks into contact points and manipulation directions. This execution interface also extends to more complex end effectors. AffordDexGrasp [187] and ScaleADFG [188] guide functional dexterous grasps, whereas DualAfford [179], BiAssemble [189], A3D [191], and BiPreManip [192] model collaborative affordances for bimanual manipulation, including assembly, coordination, and preparatory actions.

Retrieval-based transfer treats affordance as interaction knowledge that can be reused across objects. These methods retrieve a relevant exemplar, align it with the target object or part, and transfer spatial constraints, contact patterns, or motion cues for downstream interaction. DINOBot [183] retrieves a task-matched demonstration, aligns its bottleneck pose through pixel correspondences, and replays the associated trajectory. Robo-ABC [9] constructs a contact-point memory from human videos, retrieves visually or semantically similar reference objects, and transfers their contact points to novel objects through cross-category correspondences. RAM [53] broadens both the memory sources and action representation by retrieving task-relevant demonstrations from heterogeneous robot and human-interaction data, then transferring their 2D affordances into executable 3D contact points and post-contact directions. RAAP [193] further transfers contact points through dense correspondence while estimating action directions from multiple retrieved references. Together, these methods support affordance-guided action on novel objects by reusing prior interaction knowledge rather than learning a task-specific predictor from scratch.

Planner-based execution uses affordance estimates as spatial targets, objectives, or constraints for downstream planning and optimization. Rather than associating each affordance with a predefined action primitive, it lets a separate planner determine a robot-feasible viewpoint, approach motion, grasp configuration, waypoint, or trajectory. HULC++ [15] predicts a language-conditioned 3D interaction location, uses model-based planning to approach it, and then switches to a learned visuomotor policy for local execution. PreAfford [184] is a pre-grasping planner that uses point-level affordance maps to select pushing contacts and directions that improve subsequent graspability, while ACE-NBV [182] extends this paradigm to active view planning, using predicted grasp quality over imagined viewpoints to select the next view. Other methods express functional requirements as spatial fields or optimizable constraints. VoxPoser [16] composes language-derived 3D affordance and constraint maps as objectives for model-based motion planning; ReKep [17] optimizes multistage motions through relational costs over tracked 3D keypoints; and O3Afford [131] grounds object-to-object affordances to generate task-specific constraints for object-transformation planning. This principle also extends to specialized embodiments and learned action models. FSAG [194] predicts finger-specific contact affordances from human demonstrations and converts them into dexterous grasps through kinematics-aware retargeting

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

and motion planning. AFI [195] uses a 3D spatial affordance field to intervene in VLA execution through corrective waypoint generation and trajectory reranking. Across these methods, affordance specifies task-relevant spatial structure, while planning converts it into robot-feasible observation and motion decisions.

5.2. Affordance-Integrated Policy Learning

Affordance-integrated policy learning embeds affordance directly into action prediction or policy training, rather than passing it to a separate execution module. Existing approaches use affordance in three non-exclusive roles: **explicit input** directly provides grounded affordance information to the policy; **implicit representation** learns affordance-related features or auxiliary predictions within the action model; and **optimization signal** filters candidate actions, shapes rewards, or defines goals. These roles describe the primary use of affordance rather than mutually exclusive architectures: a method may, for example, condition a policy on an affordance map while also using it to shape rewards.

Affordance as explicit input directly conditions the action model on grounded affordance cues, such as interaction types, grounded regions, contact points, or motion cues:

$$\mathcal{F} = f_\theta(o, l) = (a, r, \eta), \quad u = \pi_\phi(o, l, p, \mathcal{F}). \tag{3}$$

Here, a, r, and η denote the interaction type, grounded region, and optional execution cue, while the policy π_ϕ receives $\mathcal{F}$ together with observation o, task information l, and proprioception p. IntExp [210] is an early example, learning action-specific affordance maps from interaction outcomes and using them to train an actor–critic exploration policy. VLA-based methods employ richer explicit intermediates. RT-Affordance [18] conditions action prediction on projected end-effector poses at key task stages, whereas CoA-VLA [62] represents object, grasp, spatial, and movement affordances in textual and visual forms to guide action prediction. Explicit cues also guide generative policies. Zhang and Gienger [211] condition flow-matching trajectory generation on affordance maps; AffordDP [61] conditions a diffusion policy on transferred 3D contact points and post-contact trajectories and applies affordance guidance during sampling; and DART [200] combines task representations with learned interaction points for articulated-object manipulation. Other methods focus on acquiring diverse affordance cues. UAD [49] distills task-conditioned affordance maps from large vision and vision–language models without manual annotations and uses them as observations for imitation learning. GLOVER++ [201] learns affordance representations from large-scale human-behavior data to modulate attention in policies. Affordance2Action [204] predicts scene-level, task-conditioned affordance regions as visual conditions for an action-chunking diffusion policy. Across these designs, explicit affordances remain inspectable conditions that provide functional, spatial, or motion guidance for action prediction.

Affordance as implicit representation embeds functional information within the action model rather than providing a separately predicted external input. Under *Latent affordance conditioning*, an affordance-related feature is inferred and fused with the policy inputs:

$$z_{\text{aff}} = h_\theta(o, l), \quad u = \pi_\phi(o, l, p, z_{\text{aff}}). \tag{4}$$

Alternatively, *Affordance auxiliary supervision* trains a shared model to predict both actions and affordances:

$$(u, \mathcal{F}) = \pi_\phi(o, l, p), \quad \mathcal{L} = \mathcal{L}_{\text{act}}(u, u^*) + \lambda_{\text{aff}}\mathcal{L}_{\text{aff}}(\mathcal{F}, \mathcal{F}^*), \tag{5}$$

where u^* and $\mathcal{F}^*$ denote the ground-truth action and affordance targets, respectively, and λ_{aff} balances the affordance supervision. Latent affordance conditioning introduces z_{aff} as an internal action condition, whereas auxiliary supervision uses $\mathcal{F}$ as a training signal to encourage the shared policy π_ϕ to capture task-relevant interaction structure, without requiring the affordance output at deployment.

These implicit uses of affordance are increasingly adopted in VLA-oriented policies: PALM [63] uses progress-aware affordance reasoning to track long-horizon task evolution; AffordVLA [47] distills affordance knowledge into intermediate visual representations through implicit feature alignment;

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

Afford-VLA [64] learns task-conditioned affordance regions as action-aligned visual intermediates for action generation; and AffordanceVLA [65] trains internal affordance queries through object grounding, interaction localization, and 3D geometric prediction, and allows the action expert to attend to the affordance expert's representations. These VLA methods integrate affordance learning directly with action prediction, allowing affordance-related features and objectives to be optimized jointly with the robot embodiment, data distribution, and action space.

Affordance as optimization signal directly changes the actions or outcomes favored by a policy through *action filtering*, *reward shaping*, or *goal specification*. Unlike auxiliary supervision, it modifies the feasible action set, behavioral objective, or exploration target rather than only the learned representation.

For filtering and reward shaping, Cruz et al. [212] propose a contextual-affordance-based reinforcement learning framework that predicts the effects of candidate actions and retains only those expected to produce valid states. GRAFF [197] conditions PPO on an affordance map and hand–affordance distances while penalizing distance to the predicted region. VAPO [14] learns affordance centers from teleoperated play and rewards a local SAC policy for approaching them, whereas RLAfford [198] learns point-wise affordances from RL contacts and rewards motion toward the highest-scoring point. Similarly, Nakamoto et al. [213] learn a value function with offline RL and use it to re-rank actions proposed by generalist robot policies at deployment, providing a task-conditioned affordance signal for action selection. In dexterous manipulation, DORA [202] constrains the policy with affordance-aligned grasps and rewards fingertip proximity to positive regions, while AffordDex [203] provides predicted negative-affordance regions to a residual policy and penalizes the hand for approaching them. For goal specification, VAL [196] learns an outcome model, samples latent goals, and optimizes a goal-conditioned policy with off-policy RL. Bharadhwaj et al. [199] instead sample future-image goals from passive videos and train on hindsight-relabeled trajectories through behavior cloning. Together, these methods use affordance to restrict candidate actions, reshape policy objectives, or determine which outcomes should be pursued.

6. Data and Evaluation

6.1. Datasets for Affordance Learning

Affordance datasets have expanded along several dimensions, from isolated objects to full scenes, from RGB images to videos and 3D representations, and from manually annotated masks to interaction-derived heatmaps, language-conditioned regions, contact points, and motion-related supervision. Table 4 summarizes representative datasets across these dimensions, including their setting, category and scale statistics, input modalities, and annotation types. To facilitate the following discussion, we further organize these datasets according to their primary source of affordance evidence and evaluation objective into five broad and partially overlapping groups: object-centric, scene-level, interaction-driven, language- and reasoning-oriented, and action-oriented datasets. These groups are not mutually exclusive, but provide a coherent organization for discussing the principal capability each dataset is designed to assess.

Object-centric datasets ask which part of an object supports an action. Early 2D benchmarks such as UMD [1] and IIT-AFF [3] established dense part-affordance segmentation with closed labels such as grasp, cut, contain, or support. Their 3D counterparts ground the same question in object geometry: 3D AffordanceNet [5] provides probabilistic point-wise affordance annotations and benchmarks full-shape, partial-view, and rotation-invariant estimation; PartAfford [69] contains over 25,000 objects with 24 part-affordance categories; and AffordPose [43] pairs part affordances with 3D hand poses. Together, they support actionable-part localization in both image space and metric object representations.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

Table 4. Representative affordance datasets and benchmarks.

Name	Venue & Year	Setting	Obj	Aff.	Num.	Input						Ann.			
						I	D	V	3D	L	HOI	M	H	P	Mo
UMD [1]	ICRA 2015	2D object	17	7	30k img	●	●	○	○	○	○	●	○	○	○
NYUv2 Affordance [98]	ECCV 2016	2D scene	–	5	1.4k img	●	○	○	○	○	○	●	○	○	○
IIT-AFF [3]	IROS 2017	2D object	10	9	8.8k img	●	●	○	○	○	○	●	○	○	○
OPRA [7]	CVPR 2018	2D object	–	7	2.5k img/11.5k vid	●	○	●	○	○	●	○	●	○	○
ADE-Affordance [136]	CVPR 2018	2D scene	150	3	10k img	●	○	○	○	○	○	●	○	○	○
3D AffordanceNet [5]	CVPR 2021	3D object	23	18	23k pc	○	○	○	●	○	○	○	●	○	○
PartAfford [69]	ECCVW 2022	3D object	–	24	25k pc	○	○	○	●	○	○	●	○	○	○
AGD20K [45]	CVPR 2022	2D image	50	36	23.8k img	●	○	○	○	○	●	○	●	○	○
PIADv1 [72]	ICCV 2023	3D object	23	17	5k img/7k pc	●	○	○	●	○	●	○	●	○	○
3DOI [44]	ICCV 2023	2D scene	–	3	10k img	●	○	○	○	○	○	●	○	●	●
EPIC-Aff [111]	ICCV 2023	2D/3D scene	304	20/43	38.9k img	●	○	●	○	○	●	●	○	○	○
AffordPose [43]	ICCV 2023	3D HOI	13	8	26.7k meshes	○	○	○	●	○	●	●	○	○	○
HANDAL [214]	IROS 2023	2D/3D object	17	–	308k img/2.2k vid	●	●	●	●	○	○	●	○	○	○
SceneFun3D [6]	CVPR 2024	3D scene	–	9	710 pc	○	○	○	●	●	○	●	○	○	●
3DIR [74]	CVPR 2024	3D HOI	21	17	5k img/5.1k pc	●	○	○	●	○	●	●	●	○	○
LASO [75]	CVPR 2024	3D object	23	17	8k pc	○	○	○	●	●	○	○	●	○	○
PIADv2 [147]	CVPR 2025	3D object	43	24	15k img/38.8k pc	●	○	○	●	●	●	○	●	○	○
AGPIL [146]	CVPR 2025	3D object	23	17	42k pc/31k img	●	○	○	●	●	●	○	●	○	○
SeqAfford [145]	CVPR 2025	3D object	23	18	18.4k pc	○	○	○	●	●	○	○	●	○	○
FunGraph3D [215]	CVPR 2025	3D scene	–	–	14 scenes	●	●	●	●	○	●	●	○	○	○
InstructPart [168]	ACL 2025	2D object	48	30	2.4k img	●	○	○	○	●	○	●	○	○	○
SeqAffordSplat [152]	arXiv 2025	3D scene	21	18	1.8k pc	○	○	○	●	●	○	●	○	○	○
HOVA-500K [201]	CoRL 2025	2D image	1726	675	500k img	●	○	○	○	●	○	○	○	●	○
RAGNet [169]	ICCV 2025	2D image	180	–	273k img	●	○	○	○	●	○	●	○	○	○
RoboAfford [170]	ACM MM 2025	2D scene	–	–	820k img/1.9m QA	●	○	○	○	●	○	○	○	●	○
RoboAfford++ [171]	IROSW 2025	2D scene	–	–	870k img/2.0m QA	●	○	○	○	●	○	○	○	●	○
2HANDS [216]	ICCV 2025	2D object	163	73	278k img	●	○	●	○	●	●	●	○	○	○
3DAffordSplat [154]	ACM MM 2025	3D object	21	18	8.4k pc/23.7k gs	○	○	○	●	●	○	○	●	○	○
ReasonAff [155]	AAAI 2026	2D image	48	30	2.4k img	●	○	○	○	●	○	●	○	○	○
VIDA [127]	arXiv 2026	3D object	38	16	38k vid/22k pc	○	○	●	●	●	●	○	●	○	○
360-AGD [101]	arXiv 2026	2D scene	–	19	–	●	○	○	○	●	○	○	●	○	○
CompassAD [27]	arXiv 2026	3D object	105	16	6.4k pc	○	○	○	●	●	○	○	●	○	○
A2A-Bench [204]	arXiv 2026	2D scene	116	101	15.4k img	●	○	○	○	●	○	●	○	○	○
AffordBridge [86]	CVPR 2026	3D scene	157	61	685 pc/9.8k img	●	○	○	●	○	○	●	○	●	○
Affogato-750K [88]	ECCV 2026	3D object	>450	>350	150k pc	○	○	○	●	●	○	○	●	○	○
PAP-12K [102]	ECCV 2026	2D scene	–	–	1k img	●	○	○	○	●	○	●	○	○	○

Obj. and **Aff.** denote the number of object and affordance categories when explicitly reported. **Num.** reports primary data assets. Input modalities are denoted by **I** (RGB image), **D** (depth), **V** (video), **3D** (e.g., point cloud or Gaussian representation), **L** (language/task query), and **HOI** (human–object interaction input). Annotation types are denoted by **M** (discrete mask), **H** (continuous 2D/3D heatmap), **P** (keypoint), and **Mo** (motion/trajectory).

Scene-level datasets evaluate affordances under broader environmental context. NYUv2 Affordance [98] provides dense pixel-level affordance labels for indoor scenes, whereas ADE-Affordance [136] annotates scene-dependent action–object relations together with explanations and consequences. Panoramic benchmarks such as 360-AGD [101] and PAP-12K [102] introduce wide-field observations for holistic embodied perception. In 3D settings, SceneFun3D [6] annotates functional elements and their motion parameters in full scenes, FunGraph3D [215] complements this representation with open-vocabulary 3D scene graphs that capture objects, interactive elements, and their functional relationships, and AffordBridge [86] links 2D visual signifiers to 3D affordance regions. These benchmarks extend scene evaluation from pixel-level localization to panoramic, graph-structured, and cross-modal 3D grounding.

Interaction-driven datasets infer affordances from observed human pose, contact, and motion, providing evidence of both where and how interactions occur. OPRA [7] pairs product-demonstration clips with target images, action labels, and point annotations of interaction regions, while AGD20K [45] combines exocentric interaction images with egocentric object views and point-derived heatmaps. Egocentric datasets provide complementary forms of interaction-derived supervision: EPIC-KITCHENS Affordance [217] annotates motor actions, grasp types, and mechanical interactions, whereas EPIC-Aff [111] organizes interaction-derived, multi-label annotations into metric affordance maps and spatial action-centric zones. For 3D grounding, PIAD [72] pairs interaction images with object point clouds,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

PIADv2 [147] extends this setting to open-vocabulary evaluation, and VIDA [127] associates HOI videos with point clouds and 3D affordance annotations. These resources support transferring human interaction evidence to object- and scene-level affordance grounding.

Language- and reasoning-oriented datasets condition affordance grounding on instructions, questions, or implicit goals. LLMaFF [46] pairs natural-language instructions with complex multi-object scenes, InstructPart [168] associates object part masks with task-oriented instructions, and ReasonAff [155] links affordance masks to reasoning-oriented instructions. Datasets organized primarily around question–answer evaluation include A4Bench [167], which covers constitutive and transformative reasoning, and RoboAfford++ [171], which evaluates object recognition, functional-part localization, and free-space localization. In 3D, LASO [75] pairs object point clouds with situated affordance questions and point-wise maps, CompassAD [27] uses implicit-intent queries over functionally competing objects, and Affogato-750K [88] provides open-vocabulary queries with 3D heatmaps. These datasets cover both functional reasoning and language-conditioned grounding across objects, parts, points, and regions.

Action-oriented datasets associate affordance annotations with grasping or manipulation objectives. HANDAL [214] provides 308K annotated frames of 212 real-world manipulable objects, together with category-level 6-DoF pose and scale annotations, affordance segmentation annotations, and reconstructed meshes. RAGNet [169] pairs affordance masks with template- and reasoning-based instructions for general grasping. HOVA-500K [201] derives large-scale affordance-point supervision from human behavior, and A2A-Bench [204] provides task-conditioned functional-region masks in multi-object scenes. ScaleADFG [188] evaluates functional dexterous grasping, whereas AffordSim [218] benchmarks scalable affordance-aware manipulation. Some of these datasets additionally evaluate downstream execution, whereas others remain grounding benchmarks whose annotations are designed to support action.

6.2. Evaluation Metrics

Evaluation protocols depend on the form of affordance output, spanning heatmaps, segmented regions, contact points, motion parameters, and trajectories. We therefore group commonly used metrics into heatmap/distribution, segmentation, and keypoint/motion metrics, while excluding reasoning and manipulation metrics whose definitions vary substantially across tasks and robot embodiments.

Heatmap and distribution metrics. An affordance heatmap assigns an interaction likelihood to each pixel or 3D point. Kullback–Leibler divergence (KLD) measures deviation from the target distribution, similarity (SIM) measures shared probability mass, and normalized scanpath saliency (NSS) evaluates standardized responses at positive ground-truth locations. Mean absolute error (MAE) and mean squared error (MSE) instead compare map values point-wise, with MSE placing greater weight on large deviations. Distribution metrics emphasize whether probability mass falls on the correct functional area, whereas point-wise errors measure global numerical agreement. A method may therefore achieve a low global error without sharply localizing the most actionable region, so reporting complementary metrics is generally more informative than relying on one score. Representative applications include AGD20K [45], which evaluates 2D affordance heatmaps using KLD, SIM, and NSS. Under heatmap-based evaluation in 3D, 3D AffordanceNet [5] reports point-wise MSE, whereas PIAD [72] reports SIM and MAE.

Segmentation metrics. For discrete masks or binarized maps, intersection over union (IoU) measures spatial overlap, while mean IoU averages class-wise IoU across affordance categories. In 2D evaluation, the mean of per-image IoUs, often denoted gIoU, weights each image equally, whereas cumulative IoU (cIoU) aggregates intersections and unions over the dataset and consequently gives larger regions greater influence. Thresholded precision measures the proportion of predictions attaining sufficient overlap: P@50 uses an IoU threshold of 0.5, while P@50:95 averages this success rate over thresholds from 0.50 to 0.95 in increments of 0.05. Image-level benchmarks such as InstructPart and RAGNet

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

report these region-based metrics [168,169]. Their results may differ substantially with region size, aggregation strategy, and the treatment of empty predictions.

For continuous 3D point maps, segmentation-style evaluation generally binarizes the ground-truth scores, commonly at 0.5. The area under the receiver operating characteristic curve (ROC-AUC) and mean average precision (mAP) compare continuous prediction scores against these binary targets: ROC-AUC measures ranking across decision thresholds, while mAP averages category-wise precision–recall performance and is generally more informative when positive affordance points are sparse. Averaged IoU (aIoU) instead thresholds each prediction at multiple values and averages the resulting IoUs across thresholds and samples. These metrics are commonly used by 3D affordance benchmarks [5,72], but comparisons require consistent ground-truth binarization, prediction thresholds, class averaging, and sample-level aggregation.

Keypoint and motion metrics. Keypoints are typically evaluated by Euclidean localization error or accuracy within a distance threshold. Since distances may use pixel, normalized, or metric 3D coordinates, the coordinate system and scale should be reported. Motion-aware protocols additionally evaluate interaction dynamics. SceneFun3D [6], for example, augments mask average precision with requirements on motion type, axis orientation, and axis origin, separating localization from motion-estimation errors.

For trajectories, average displacement error (ADE) compares temporally corresponding positions, whereas dynamic time warping (DTW) allows nonlinear alignment and is less sensitive to execution speed. Lower ADE and DTW values indicate more accurate trajectories. Evaluation protocols may also normalize trajectories before computing these metrics. For example, Yoshida et al. [117] normalize predicted and ground-truth trajectories and set their initial coordinates to zero, isolating trajectory shape from contact-point localization.

7. Challenges and Future Directions

Affordance learning has evolved from visual recognition toward a bridge between perception, reasoning, and action in embodied AI. Despite this progress, important challenges remain in developing scalable, reliable, and actionable affordance learning systems for open-world embodied interaction. Accordingly, we organize the following discussion around affordance representation and action integration, scalable learning and open-world generalization, reliable and autonomous grounding, closed-loop modeling and planning, and embodied evaluation and trustworthiness.

7.1. Affordance Representation and Action Integration

A central question is not only what an affordance representation should predict, but how it should contribute to robot action. Explicit masks, contacts, or trajectories provide interpretable interfaces for planners and policies, but make execution sensitive to prediction errors [9,18,62,118]. Latent features enable flexible end-to-end optimization but often require additional feature encoders and provide limited interpretability [47,65]. Auxiliary objectives introduce action-oriented inductive bias without requiring affordance outputs at deployment, yet improved auxiliary prediction does not necessarily yield higher execution success [64,65]. Since these designs are usually evaluated with different data, architectures, action spaces, and tasks, it remains unclear when each form of integration is most effective.

Future work should compare explicit, latent, auxiliary, optimization-signal, and hybrid formulations through controlled downstream experiments with matched observations, training data, policy backbones, and action interfaces. Comparisons should include no-affordance, predicted-affordance, and oracle-affordance conditions rather than rely on localization performance alone. Intervention tests that remove, perturb, or replace affordance information can further reveal whether a policy causally depends on the intended functional signal. The broader goal is to identify not one universally optimal interface, but the conditions under which each affordance representation most effectively improves action.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

7.2. Scalable Learning and Open-World Generalization

Dense affordance annotation is difficult to scale because annotators must infer fine-grained, action-specific interaction regions. Existing methods therefore draw supervision from human–object interactions, demonstrations, pseudo-labels from foundation models, simulation, and robot interaction [7,8,18,49,60,71,219]. These sources broaden coverage but introduce different forms of uncertainty and bias: human bodies may occlude contacts, videos may suffer from motion blur and limited visibility of contact regions, and pseudo-labels may inherit foundation-model hallucinations and semantic preferences. Moreover, many datasets remain concentrated on a small set of object categories, canonical actions, clean observations, and limited object states. Models can consequently perform well in an in-distribution setting while failing to transfer to unseen objects, novel tasks, or complex environments. Scalability and open-world generalization are therefore inseparable: additional data is valuable only when its diversity and quality support transferable functional relations.

Future data pipelines should expand coverage while actively controlling supervision quality. Affordance-guided data generation offers a promising direction by preserving functional correspondences while varying object geometry, pose, and scene configuration. Recent work instantiates this idea by transferring limited demonstrations across novel 3D objects or integrating open-vocabulary prediction with simulation to generate manipulation data [218,220,221]. Generative predictive models may further scale affordance learning by synthesizing diverse interaction trajectories and state transitions [222,223]. The objective is not merely to enlarge the training set but to learn functional relations that remain stable across objects, tasks, states, and environments.

7.3. Reliable and Autonomous Affordance Grounding

Foundation models have transformed affordance grounding from single-model prediction into multi-stage pipelines. A typical workflow may use a vision-language model to interpret the instruction and identify a relevant object or semantic part, a detector or segmentation foundation model to localize that part, and additional 3D, geometric, or verification modules to obtain the final interaction region [49,224]. Modular and agentic systems make these stages explicit and can combine complementary capabilities that no single model provides [102,163]. However, their performance often depends on manually chosen prompts, part vocabularies, model combinations, thresholds, views, execution orders, and fallback rules. A segmentation model may recover visual boundaries rather than the functional contact area, and an early grounding error may propagate through every subsequent stage. This makes it difficult to determine whether gains arise from a general affordance capability or from task-specific pipeline engineering, and complicates reproducibility and fair comparison across rapidly changing foundation models.

Future systems should move from manually assembled feed-forward pipelines toward automatically orchestrated and self-verifying grounding processes. An autonomous grounding system should select tools and evidence sources according to the task, preserve the provenance and uncertainty of intermediate outputs, check semantic, spatial, cross-view, and cross-module consistency, and trigger targeted retries or additional evidence acquisition only when necessary. Adaptive routing and affordance-specific verification provide initial steps in this direction [160,161], but systematic evaluation is still needed. In addition to final grounding accuracy, comparisons should report intermediate failure modes, the number of human corrections, prompt and rule dependence, tool calls, latency, computational cost, and sensitivity to component replacement. The aim is not simply to construct a longer pipeline, but to reduce manual intervention while making its decisions traceable and robust to component errors.

7.4. Closed-Loop Affordance Modeling and Planning

Affordances are state-dependent rather than fixed labels. Robot actions can create occlusions, alter object poses or scene configurations, and invalidate previously correct estimates. In long-horizon tasks [225,226], each outcome also changes which object, region, or interaction is relevant next. Nev-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

ertheless, most methods rely on one-shot prediction, allowing early errors to compound without closed-loop correction or continual affordance updating.

Future systems should maintain affordance as a belief updated through observation, action, and execution feedback. Active observation and low-risk exploratory actions can reveal hidden articulation, friction, stability, compliance, and occluded parts, while uncertainty and expected information gain can determine when an agent should gather evidence, act, or revise its hypothesis [160,178,182,227]. For long-horizon tasks, agents must also model affordance transitions, use execution outcomes to revise later choices, track progress, and recover from unexpected effects. Sequential reasoning, relational keypoint constraints, and multi-stage scene representations provide foundations for this direction [17,132,145,152,228]. This would turn affordance into a persistent interface connecting perception, reasoning, action, and feedback over time.

7.5. Embodied Evaluation, Safety, and Trustworthiness

Common metrics such as classification accuracy, mask overlap, point localization, and heatmap similarity measure perceptual agreement but not embodiment-aware validity. A plausible region may be unreachable, unstable, or incompatible with the agent's morphology, kinematics, or skills, making validity inherently agent-dependent. Recent systems such as SceneTeract [159] explicitly model agent profiles and feasibility, but most benchmarks still overlook embodiment-specific executability. Safety further separates what an agent *can* do from what it *should* do: negative affordances, fragile regions, social constraints, and embodiment-specific failures remain underexplored.

Future benchmarks should evaluate affordance as an embodied and trustworthy capability, complementing localization with embodiment-specific feasibility, downstream success, and uncertainty calibration across tasks and robot platforms, with validated simulation providing a scalable evaluation complement [229]. Building on multisensory policy learning [230,231], evaluation should also test whether integrating complementary modalities improves affordance judgments under ambiguous visual evidence. Systems should condition validity judgments on the agent, act conservatively under uncertainty, and account for human intent and social norms [232–234], since physical executability alone does not ensure safe or appropriate interaction.

References

1. Myers, A.; Teo, C.L.; Fermüller, C.; Aloimonos, Y. Affordance detection of tool parts from geometric features. In Proceedings of the ICRA, 2015, pp. 1374–1381.
2. Nguyen, A.; Kanoulas, D.; Caldwell, D.G.; Tsagarakis, N.G. Detecting object affordances with convolutional neural networks. In Proceedings of the IROS, 2016, pp. 2765–2770.
3. Nguyen, A.; Kanoulas, D.; Caldwell, D.G.; Tsagarakis, N.G. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In Proceedings of the IROS, 2017, pp. 5908–5915.
4. Do, T.T.; Nguyen, A.; Reid, I. Affordancenet: An end-to-end deep learning approach for object affordance detection. In Proceedings of the ICRA, 2018, pp. 5882–5889.
5. Deng, S.; Xu, X.; Wu, C.; Chen, K.; Jia, K. 3d affordancenet: A benchmark for visual object affordance understanding. In Proceedings of the CVPR, 2021, pp. 1778–1787.
6. Delitzas, A.; Takmaz, A.; Tombari, F.; Sumner, R.; Pollefeys, M.; Engelmann, F. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In Proceedings of the CVPR, 2024, pp. 14531–14542.
7. Fang, K.; Wu, T.L.; Yang, D.; Savarese, S.; Lim, J.J. Demo2vec: Reasoning object affordances from online videos. In Proceedings of the CVPR, 2018, pp. 2139–2147.
8. Liu, S.; Tripathi, S.; Majumdar, S.; Wang, X. Joint hand motion and interaction hotspots prediction from egocentric videos. In Proceedings of the CVPR, 2022, pp. 3282–3292.
9. Ju, Y.; Hu, K.; Zhang, G.; Zhang, G.; Jiang, M.; Xu, H. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In Proceedings of the ECCV, 2024, pp. 222–239.
10. Bahl, S.; Mendonca, R.; Chen, L.; Jain, U.; Pathak, D. Affordances from human videos as a versatile representation for robotics. In Proceedings of the CVPR, 2023, pp. 13778–13790.
11. Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv:2204.01691* **2022**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

12. Huang, W.; Xia, F.; Shah, D.; Driess, D.; Zeng, A.; Lu, Y.; Florence, P.; Mordatch, I.; Levine, S.; Hausman, K.; et al. Grounded decoding: Guiding text generation with grounded models for embodied agents. *NeurIPS* **2023**, *36*, 59636–59661.
13. Qian, S.; Chen, W.; Bai, M.; Zhou, X.; Tu, Z.; Li, L.E. AffordanceLLM: Grounding Affordance from Vision Language Models. In Proceedings of the CVPRW, June 2024, pp. 7587–7597.
14. Borja-Diaz, J.; Mees, O.; Kalweit, G.; Hermann, L.; Boedecker, J.; Burgard, W. Affordance learning from play for sample-efficient policy learning. In Proceedings of the ICRA, 2022, pp. 6372–6378.
15. Mees, O.; Borja-Diaz, J.; Burgard, W. Grounding Language with Visual Affordances over Unstructured Data. In Proceedings of the ICRA, London, UK, 2023.
16. Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; Fei-Fei, L. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In Proceedings of the CoRL, 2023.
17. Huang, W.; Wang, C.; Li, Y.; Zhang, R.; Fei-Fei, L. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv:2409.01652* **2024**.
18. Nasiriany, S.; Kirmani, S.; Ding, T.; Smith, L.; Zhu, Y.; Driess, D.; Sadigh, D.; Xiao, T. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. In Proceedings of the ICRA, 2025, pp. 8249–8257.
19. Hassanin, M.; Khan, S.; Tahtali, M. Visual affordance and function understanding: A survey. *ACM CSUR* **2021**, *54*, 1–35.
20. Chen, D.; Kong, D.; Li, J.; Wang, S.; Yin, B. A survey of visual affordance recognition based on deep learning. *IEEE Transactions on Big Data* **2023**, *9*, 1458–1476.
21. Apicella, T.; Xompero, A.; Cavallaro, A. Visual Affordance Prediction: Survey and Reproducibility. *arXiv:2505.05074* **2025**.
22. Jamone, L.; Ugur, E.; Cangelosi, A.; Fadiga, L.; Bernardino, A.; Piater, J.; Santos-Victor, J. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE TCDS* **2016**, *10*, 4–25.
23. Min, H.; Yi, C.; Luo, R.; Zhu, J.; Bi, S. Affordance research in developmental robotics: A survey. *IEEE TCDS* **2016**, *8*, 237–255.
24. Ardón, P.; Pairet, È.; Lohan, K.S.; Ramamoorthy, S.; Petrick, R. Affordances in robotic tasks–a survey. *arXiv:2004.07400* **2020**.
25. Yang, X.; Ji, Z.; Wu, J.; Lai, Y.K. Recent advances of deep robotic affordance learning: A reinforcement learning perspective. *IEEE TCDS* **2023**.
26. Yu, Q.; Huang, S.; Yuan, X.; Jiang, Z.; Hao, C.; Li, X.; Chang, H.; Wang, J.; Liu, L.; Li, H.; et al. Uniaff: A unified representation of affordances for tool usage and articulation with vision-language models. In Proceedings of the ICRA, 2025, pp. 8980–8987.
27. Li, J.; Jia, J.; An, T.; Zhou, C.; Chen, X.; Shan, S.; Ma, B.; Lyu, B.; Li, G.; Yang, J. CompassAD: Intent-Driven 3D Affordance Grounding in Functionally Competing Objects. *arXiv:2604.02060* **2026**.
28. Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv:2307.15818* **2023**.
29. Kim, M.J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. Openvla: An open-source vision-language-action model. *arXiv:2406.09246* **2024**.
30. Octo Model Team.; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; et al. Octo: An Open-Source Generalist Robot Policy. In Proceedings of the RSS, Delft, Netherlands, 2024.
31. Doshi, R.; Walke, H.; Mees, O.; Dasari, S.; Levine, S. Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation. In Proceedings of the CoRL, 2024.
32. Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; Levine, S. FAST: Efficient Action Tokenization for Vision-Language-Action Models. In Proceedings of the RSS, Los Angeles, USA, 2025.
33. Zawalski, M.; Chen, W.; Pertsch, K.; Mees, O.; Finn, C.; Levine, S. Robotic Control via Embodied Chain-of-Thought Reasoning. In Proceedings of the CoRL, 2024.
34. Pai, J.; Achenbach, L.; Montesinos, V.; Forrai, B.; Mees, O.; Nava, E. mimic-video: Video-Action Models for Generalizable Robot Control Beyond VLAs. In Proceedings of the RSS, Sydney, Australia, 2026.
35. Zhou, H.; Yao, X.; Mees, O.; Meng, Y.; Xiao, T.; Bisk, Y.; Oh, J.; Johns, E.; Shridhar, M.; Shah, D.; et al. Bridging Language and Action: A Survey of Language-Based Robot Manipulation. *IJRR* **2026**.
36. Gibson, J.J. *The Ecological Approach to Visual Perception*; Houghton Mifflin: Boston, 1979.
37. Bajcsy, R. Active perception. *Proceedings of the IEEE* **1988**, *76*, 966–1005.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

38. Bogoni, L.; Bajcsy, R. Interactive recognition and representation of functionality. *CVIU* **1995**, *62*, 194–214.
39. Stoytchev, A. Behavior-grounded representation of tool affordances. In Proceedings of the ICRA, 2005, pp. 3060–3065.
40. Fritz, G.; Paletta, L.; Breithaupt, R.; Rome, E.; Dorffner, G. Learning Predictive Features in Affordance Based Robotic Perception Systems. In Proceedings of the IROS, 2006, pp. 3642–3647.
41. Stark, M.; Lies, P.; Zillich, M.; Wyatt, J.; Schiele, B. Functional object class detection based on learned affordance cues. In Proceedings of the ICVS, 2008, pp. 435–444.
42. Nagarajan, T.; Feichtenhofer, C.; Grauman, K. Grounded human-object interaction hotspots from video. In Proceedings of the ICCV, 2019, pp. 8688–8697.
43. Jian, J.; Liu, X.; Li, M.; Hu, R.; Liu, J. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In Proceedings of the ICCV, 2023, pp. 14713–14724.
44. Qian, S.; Fouhey, D.F. Understanding 3d object interaction from a single image. In Proceedings of the ICCV, 2023, pp. 21753–21763.
45. Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; Tao, D. Learning affordance grounding from exocentric images. In Proceedings of the CVPR, 2022, pp. 2252–2261.
46. Chen, C.; Cong, Y.; Kan, Z. Worldafford: Affordance grounding based on natural language instructions. In Proceedings of the ICTAI, 2024, pp. 822–828.
47. Kong, W.; Su, Z.; Yu, W.; Dong, H. AffordVLA: Injecting Affordance Representations into Vision-Language-Action Models via Implicit Feature Alignment. *arXiv:2605.17517* **2026**.
48. Sawatzky, J.; Srikantha, A.; Gall, J. Weakly supervised affordance detection. In Proceedings of the CVPR, 2017, pp. 2795–2804.
49. Tang, Y.; Huang, W.; Wang, Y.; Li, C.; Yuan, R.; Zhang, R.; Wu, J.; Fei-Fei, L. Uad: Unsupervised affordance distillation for generalization in robotic manipulation. In Proceedings of the ICRA, 2025, pp. 3822–3831.
50. Mo, K.; Qin, Y.; Xiang, F.; Su, H.; Guibas, L. O2O-Afford: Annotation-free large-scale object-object affordance learning. In Proceedings of the CoRL, 2022, pp. 1666–1677.
51. Wu, H.; Misra, D.; Chirikjian, G.S. Is that a chair? imagining affordances using simulations of an articulated human body. In Proceedings of the ICRA, 2020, pp. 7240–7246.
52. Wu, H.; Chirikjian, G.S. Can i pour into it? robot imagining open containability affordance of previously unseen objects via physical simulations. *IEEE RA-L* **2020**, *6*, 271–278.
53. Kuang, Y.; Ye, J.; Geng, H.; Mao, J.; Deng, C.; Guibas, L.; Wang, H.; Wang, Y. RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation. In Proceedings of the CoRL, 2024.
54. Wang, X.; Yang, X.; Xu, Y.; Wu, Y.; Li, Z.; Zhao, N. AffordBot: 3D Fine-grained Embodied Reasoning via Multimodal Large Language Models. In Proceedings of the NeurIPS, 2025.
55. Rashid, A.; Sharma, S.; Kim, C.M.; Kerr, J.; Chen, L.Y.; Kanazawa, A.; Goldberg, K. Language embedded radiance fields for zero-shot task-oriented grasping. In Proceedings of the CoRL, 2023.
56. Wang, Z.; Zhong, Y.; Fu, J.; Christensen, H.I.; Gao, J. AFUN: Towards an Affordance Foundation Model for Functionality Understanding. *arXiv:2606.02551* **2026**.
57. Nguyen, T.; Vu, M.N.; Vuong, A.; Nguyen, D.; Vo, T.; Le, N.; Nguyen, A. Open-vocabulary affordance detection in 3d point clouds. In Proceedings of the IROS, 2023, pp. 5692–5698.
58. Nguyen, T.; Vu, M.N.; Huang, B.; Van Vo, T.; Truong, V.; Le, N.; Vo, T.; Le, B.; Nguyen, A. Language-conditioned affordance-pose detection in 3D point clouds. In Proceedings of the ICRA, 2024, pp. 3071–3078.
59. Van Vo, T.; Vu, M.N.; Huang, B.; Nguyen, T.; Le, N.; Vo, T.; Nguyen, A. Open-vocabulary affordance detection using knowledge distillation and text-point correlation. In Proceedings of the ICRA, 2024, pp. 13968–13975.
60. Yuan, W.; Duan, J.; Blukis, V.; Pumacay, W.; Krishna, R.; Murali, A.; Mousavian, A.; Fox, D. RoboPoint: A Vision-Language Model for Spatial Affordance Prediction in Robotics. In Proceedings of the CoRL, 2024.
61. Wu, S.; Zhu, Y.; Huang, Y.; Zhu, K.; Gu, J.; Yu, J.; Shi, Y.; Wang, J. Afforddp: Generalizable diffusion policy with transferable affordance. In Proceedings of the CVPR, 2025, pp. 6971–6980.
62. Li, J.; Zhu, Y.; Tang, Z.; Wen, J.; Zhu, M.; Liu, X.; Li, C.; Cheng, R.; Peng, Y.; Peng, Y.; et al. Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance. In Proceedings of the ICCV, 2025, pp. 9759–9769.
63. Liu, Y.; Zhu, J.; Mo, Y.; Li, G.; Cao, X.; Jin, J.; Shen, Y.; Li, Z.; Yu, T.; Yuan, W.; et al. PALM: Progress-Aware Policy Learning via Affordance Reasoning for Long-Horizon Robotic Manipulation. In Proceedings of the CVPR, 2026.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

64. Wang, R.; Fu, Y.; Li, Y.; Lin, T.; Qian, T.; Elhoseiny, M.; Zhao, B.; Fu, Y.; Jiang, Y.G.; Xue, X. Afford-VLA: Action-Aligned Visual Planning via Internalized Affordance. *arXiv:2605.24203* **2026**.
65. Yu, Q.; You, J.; Wang, Y.; Liang, J.; Ping, B.; Tian, Y.; Chen, Y.; Cai, M.; Gong, Z.; Wu, R.; et al. AffordanceVLA: A Vision-Language-Action Model Empowering Action Generation through Affordance-Aware Understanding. *arXiv:2606.06155* **2026**.
66. Nagarajan, T.; Li, Y.; Feichtenhofer, C.; Grauman, K. Ego-topo: Environment affordances from egocentric video. In Proceedings of the CVPR, 2020, pp. 163–172.
67. Xu, R.; Chu, F.J.; Tang, C.; Liu, W.; Vela, P.A. An affordance keypoint detection network for robot manipulation. *IEEE RA-L* **2021**, *6*, 2870–2877.
68. Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; Tao, D. One-Shot Affordance Detection. In Proceedings of the IJCAI, 2021, pp. 895–901.
69. Xu, C.; Chen, Y.; Wang, H.; Zhu, S.C.; Zhu, Y.; Huang, S. Partafford: Part-level affordance discovery from 3d objects. *arXiv:2202.13519* **2022**.
70. Hadjivelichkov, D.; Zwane, S.; Agapito, L.; Deisenroth, M.P.; Kanoulas, D. One-shot transfer of affordance regions? affcorrs! In Proceedings of the CoRL, 2023, pp. 550–560.
71. Li, G.; Jampani, V.; Sun, D.; Sevilla-Lara, L. LOCATE: Localize and Transfer Object Parts for Weakly Supervised Affordance Grounding. In Proceedings of the CVPR, 2023, pp. 10922–10931.
72. 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 ICCV, 2023, pp. 10905–10915.
73. Ning, C.; Wu, R.; Lu, H.; Mo, K.; Dong, H. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *NeurIPS* **2023**, *36*, 4585–4596.
74. Yang, Y.; Zhai, W.; Luo, H.; Cao, Y.; Zha, Z.J. Lemon: Learning 3d human-object interaction relation from 2d images. In Proceedings of the CVPR, 2024, pp. 16284–16295.
75. Li, Y.; Zhao, N.; Xiao, J.; Feng, C.; Wang, X.; Chua, T.s. Laso: Language-guided affordance segmentation on 3d object. In Proceedings of the CVPR, 2024, pp. 14251–14260.
76. Li, G.; Sun, D.; Sevilla-Lara, L.; Jampani, V. One-shot open affordance learning with foundation models. In Proceedings of the CVPR, 2024, pp. 3086–3096.
77. Liu, Z.; Zhao, R.; Zhou, L.; Yuan, C.; Wu, Y.; Guo, S.; Zhang, Z.; Liu, C.; Ang, M.H.; Tay, F.E. 3D Affordance keypoint detection for robotic manipulation. In Proceedings of the IROS, 2024, pp. 7528–7534.
78. Jang, J.H.; Seo, H.; Chun, S.Y. INTRA: Interaction Relationship-Aware Weakly Supervised Affordance Grounding. In Proceedings of the ECCV, 2024, pp. 18–34.
79. Chen, D.; Kong, D.; Li, J.; Yin, B. MaskPrompt: Open-vocabulary affordance segmentation with object shape mask prompts. In Proceedings of the AAAI, 2025, Vol. 39, pp. 2034–2042.
80. Wang, Y.; Wu, A.; Yang, M.; Min, Y.; Zhu, Y.; Deng, C. Reasoning mamba: Hypergraph-guided region relation calculating for weakly supervised affordance grounding. In Proceedings of the CVPR, 2025, pp. 27618–27627.
81. Lu, D.; Kong, L.; Huang, T.; Lee, G.H. Geal: Generalizable 3d affordance learning with cross-modal consistency. In Proceedings of the CVPR, 2025, pp. 1680–1690.
82. Li, Y.; Chen, Y.; Ma, Z.; Xiao, J.; Wang, X.; Yao, A. Intermediate connectors and geometric priors for language-guided affordance segmentation on unseen object categories. In Proceedings of the ICCV, 2025, pp. 22836–22845.
83. Tang, J.; Wei, Z.; Zheng, G.; Yang, S. Closed-Loop Transfer for Weakly-supervised Affordance Grounding. In Proceedings of the ICCV, 2025, pp. 9530–9539.
84. Delitzas, A.; Zhang, C.; Gavryushin, A.; Di Mario, T.; Sun, B.; Dabral, R.; Guibas, L.; Theobalt, C.; Pollefeys, M.; Engelmann, F.; et al. FunRec: Reconstructing Functional 3D Scenes from Egocentric Interaction Videos. In Proceedings of the CVPR, 2026.
85. Park, C.; Lee, S.; Cho, M. Affostruction: 3d affordance grounding with generative reconstruction. *arXiv:2601.09211* **2026**.
86. Vu, N.; Do, T.; Nguyen, K.; Huang, B.; Le, N.; Nguyen, B.X.; Tjiputra, E.; Tran, Q.D.; Prakash, R.; Chiu, T.C.; et al. AffordMatcher: Affordance Learning in 3D Scenes from Visual Signifiers. In Proceedings of the CVPR, 2026, pp. 2635–2644.
87. Wang, H.; Zhang, Z.; Ji, K.; Liu, M.; Yin, W.; Chen, Y.; Liu, Z.; Zeng, X.; Gui, T.; Zhang, H.; et al. Diffusion Models are Open-World Affordance Learners: Leveraging Generative Priors for 3D Affordance Learning. In Proceedings of the ECCV, 2026.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

88. Lee, J.; Park, E.; Park, C.; Kang, D.; Cho, M. Affogato: Learning Open-Vocabulary Affordance Grounding with Automated Data Generation at Scale. *arXiv:2506.12009* **2025**.
89. Jiang, D.; Wang, Z.; Li, H.; Dang, S.; Ma, T.; Wei, W.; Dai, G.; Zhang, L.; Wang, M. AffordanceSAM: Segment Anything Once More in Affordance Grounding. In Proceedings of the ACM MM, 2026.
90. Rezapour Lakani, S.; Rodríguez-Sánchez, A.J.; Piater, J. Towards affordance detection for robot manipulation using affordance for parts and parts for affordance. *Autonomous Robots* **2019**, *43*, 1155–1172.
91. Zhao, X.; Cao, Y.; Kang, Y. Object affordance detection with relationship-aware network. *Neural Computing and Applications* **2020**, *32*, 14321–14333.
92. Gu, Q.; Su, J.; Yuan, L. Visual affordance detection using an efficient attention convolutional neural network. *Neurocomputing* **2021**, *440*, 36–44.
93. Yin, C.; Zhang, Q. Object affordance detection with boundary-preserving network for robotic manipulation tasks. *Neural Computing and Applications* **2022**, *34*, 17963–17980.
94. Chu, F.J.; Xu, R.; Tang, C.; Vela, P.A. Recognizing object affordances to support scene reasoning for manipulation tasks. *arXiv:1909.05770* **2019**.
95. Chu, F.J.; Xu, R.; Vela, P.A. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE RA-L* **2019**, *4*, 1140–1147.
96. Khalifa, Z.; Shah, S.A.A. A large scale multi-view RGBD visual affordance learning dataset. In Proceedings of the ICIP, 2023, pp. 1325–1329.
97. Shah, S.A.A.; Khalifa, Z. Hierarchical transformer for visual affordance understanding using a large-scale dataset. In Proceedings of the IROS, 2023, pp. 11371–11376.
98. Roy, A.; Todorovic, S. A multi-scale cnn for affordance segmentation in rgb images. In Proceedings of the ECCV, 2016, pp. 186–201.
99. Luddecke, T.; Worgotter, F. Learning to segment affordances. In Proceedings of the ICCVW, 2017, pp. 769–776.
100. Lueddecke, T.; Kulvicius, T.; Woergoetter, F. Context-based affordance segmentation from 2D images for robot actions. *Robotics and Autonomous Systems* **2019**, *119*, 92–107.
101. Zhu, G.; Jia, W.; Shao, C.; Zhang, Y.; Li, Z.; Yang, K. PanoAffordanceNet: Towards Holistic Affordance Grounding in 360° Indoor Environments. *arXiv:2603.09760* **2026**.
102. Zhang, Z.; Liao, C.; Zhang, H.; Chen, H.H.; Chen, K.; Wen, Z.; Guo, L.; Ren, B.; Zheng, X.; Li, Y.; et al. Panoramic Affordance Prediction. *arXiv:2603.15558* **2026**.
103. Sawatzky, J.; Gall, J. Adaptive binarization for weakly supervised affordance segmentation. In Proceedings of the ICCVW, 2017, pp. 1383–1391.
104. Cui, L.; Chen, X.; Zhao, H.; Zhou, G.; Zhu, Y. STRAP: Structured Object Affordance Segmentation with Point Supervision. *arXiv:2304.08492* **2023**.
105. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the CVPR, 2016, pp. 2921–2929.
106. Kokic, M.; Stork, J.A.; Haustein, J.A.; Kragic, D. Affordance detection for task-specific grasping using deep learning. In Proceedings of the Humanoids, 2017, pp. 91–98.
107. Tabib, R.A.; Hegde, D.; Mudenagudi, U. LGAfford-Net: A Local Geometry Aware Affordance Detection Network for 3D Point Clouds. In Proceedings of the CVPRW, 2024, pp. 5261–5270.
108. Li, M.X.; Rudolf, K.; Blank, N.; Lioutikov, R. Interpretable affordance detection on 3d point clouds with probabilistic prototypes. *arXiv:2504.18355* **2025**.
109. Chen, X.; Zheng, K.; Zeng, Z.; Kisailus, C.; Basu, S.; Cooney, J.; Pavlasek, J.; Jenkins, O.C. Manipulation-oriented object perception in clutter through affordance coordinate frames. In Proceedings of the Humanoids, 2022, pp. 186–193.
110. Kim, H.; Han, S.; Kwon, P.; Joo, H. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In Proceedings of the ECCV, 2024, pp. 400–419.
111. Mur-Labadia, L.; Guerrero, J.J.; Martinez-Cantin, R. Multi-label affordance mapping from egocentric vision. In Proceedings of the ICCV, 2023, pp. 5238–5249.
112. Zhang, C.; Lee, G.H. IAAO: Interactive Affordance Learning for Articulated Objects in 3D Environments. In Proceedings of the CVPR, 2025, pp. 12132–12142.
113. Koppula, H.S.; Gupta, R.; Saxena, A. Learning human activities and object affordances from rgb-d videos. *IJRR* **2013**, *32*, 951–970.
114. Damen, D.; Doughty, H.; Farinella, G.M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV* **2022**, pp. 1–23.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

115. Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the CVPR, 2022, pp. 18995–19012.
116. Chen, J.; Gao, D.; Lin, K.Q.; Shou, M.Z. Affordance grounding from demonstration video to target image. In Proceedings of the CVPR, 2023, pp. 6799–6808.
117. Yoshida, T.; Kurita, S.; Nishimura, T.; Mori, S. Text-driven affordance learning from egocentric vision. *Adv. Robot.* **2025**, *39*, 1041–1052.
118. Li, G.; Tsagkas, N.; Song, J.; Mon-Williams, R.; Vijayakumar, S.; Shao, K.; Sevilla-Lara, L. Learning precise affordances from egocentric videos for robotic manipulation. In Proceedings of the ICCV, 2025, pp. 10581–10591.
119. Chen, H.; Zhang, A.; Schaefer, S.; Chen, K.; Chen, S.; Cremers, D.; Mees, O.; Leutenegger, S. Robot Self-Improvement via Human-Video Dynamics Models. *arXiv:2606.21406* **2026**.
120. Feng, Z.; Li, Q.; Liang, H.; Yang, R.; Shen, Y.; Du, Z.; Zhang, Z.; Deng, Y.; Zhao, L.; Zhao, H.; et al. From Human Videos to Robot Manipulation: A Survey on Scalable Vision-Language-Action Learning with Human-centric Data. In Proceedings of the IJCAI, Bremen, Germany, 2026.
121. Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; Tao, D. Leverage Interactive Affinity for Affordance Learning. In Proceedings of the CVPR, 2023, pp. 6809–6819.
122. Xu, L.; Gao, Y.; Song, W.; Hao, A. Weakly supervised multimodal affordance grounding for egocentric images. In Proceedings of the AAAI, 2024, pp. 6324–6332.
123. Huang, Y.; Yang, F.; Zhu, G.; Li, G.; Shi, H.; Zuo, Y.; Chen, W.; Li, Z.; Yang, K. Resource-Efficient Affordance Grounding with Complementary Depth and Semantic Prompts. In Proceedings of the IROS, 2025, pp. 7788–7795.
124. Xu, P.; Mu, Y. Weakly-supervised affordance grounding guided by part-level semantic priors. *arXiv:2505.24103* **2025**.
125. Moon, W.; Seong, H.S.; Heo, J.P. Selective Contrastive Learning for Weakly Supervised Affordance Grounding. In Proceedings of the ICCV, 2025, pp. 5210–5220.
126. 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, 2025, Vol. 39, pp. 3095–3103.
127. 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:2602.09638* **2026**.
128. Liu, C.; Zhai, W.; Yang, Y.; Luo, H.; Liang, S.; Cao, Y.; Zha, Z.J. Grounding 3d scene affordance from egocentric interactions. *arXiv:2409.19650* **2024**.
129. Zhai, W.; Luo, H.; Zhang, J.; Cao, Y.; Tao, D. One-shot object affordance detection in the wild. *IJCV* **2022**, *130*, 2472–2500.
130. Jia, W.; Yang, F.; Duan, M.; Chen, X.; Wang, Y.; Jiang, Y.; Chen, W.; Yang, K.; Li, Z. One-shot affordance grounding of deformable objects in egocentric organizing scenes. In Proceedings of the IROS, 2025, pp. 21243–21250.
131. Tian, T.; Kang, X.; Kuo, Y.L. O3Afford: One-Shot 3D Object-to-Object Affordance Grounding for Generalizable Robotic Manipulation. In Proceedings of the CoRL, 2025.
132. Kim, H.; Song, Y.J.; Lee, M.; Zhang, B.T. One-Shot 3-D Affordance Learning for Multi-Stage Robotic Manipulation. *IEEE Access* **2026**, *14*, 54571–54582.
133. Su, H.; Xie, M.; Cao, N.; Ding, Y.; Shao, B.; Long, X.; Gu, F.; Chen, C. OVA-Fields: Weakly Supervised Open-Vocabulary Affordance Fields for Robot Operational Part Detection. In Proceedings of the ICCV, 2025, pp. 6385–6395.
134. Zhu, Y.; Fathi, A.; Fei-Fei, L. Reasoning about Object Affordances in a Knowledge Base Representation. In Proceedings of the Computer Vision – ECCV 2014. Springer, 2014, pp. 408–424.
135. Fichtl, S.; Kraft, D.; Krüger, N.; Guerin, F. Bootstrapping relational affordances of object pairs using transfer. *IEEE TCDS* **2016**, *10*, 56–71.
136. Chuang, C.Y.; Li, J.; Torralba, A.; Fidler, S. Learning to act properly: Predicting and explaining affordances from images. In Proceedings of the CVPR, 2018, pp. 975–983.
137. Ardón, P.; Pairet, E.; Petrick, R.P.; Ramamoorthy, S.; Lohan, K.S. Learning grasp affordance reasoning through semantic relations. *IEEE RA-L* **2019**, *4*, 4571–4578.
138. Mees, O.; Emek, A.; Vertens, J.; Burgard, W. Learning Object Placements For Relational Instructions by Hallucinating Scene Representations. In Proceedings of the ICRA, Paris, France, 2020.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

139. Tang, C.; Yu, J.; Chen, W.; Zhang, H. Relationship Oriented Affordance Learning through Manipulation Graph Construction. *arXiv:2110.14137* **2021**.
140. Xin, J.; Wang, L.; Wang, S.; Kong, D.; Li, J.; Yin, B. A Visual Affordance Reasoning Network Based on Graph Attention. In Proceedings of the ICDH, 2022, pp. 283–290.
141. Huang, S.; Ponomarenko, I.; Jiang, Z.; Li, X.; Hu, X.; Gao, P.; Li, H.; Dong, H. ManipVQA: Injecting Robotic Affordance and Physically Grounded Information into Multi-Modal Large Language Models. In Proceedings of the IROS, 2024.
142. Tong, E.; Opipari, A.; Lewis, S.; Zeng, Z.; Jenkins, O.C. OVAL-Prompt: Open-Vocabulary Affordance Localization for Robot Manipulation through LLM Affordance-Grounding. *arXiv:2404.11000* **2024**.
143. 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. *ICLR* **2025**.
144. Zhu, X.; Li, Y.; Cui, L.; Li, P.; Gao, H.a.; Zhu, Y.; Zhao, H. Afford-X: Generalizable and Slim Affordance Reasoning for Task-oriented Manipulation. *arXiv:2503.03556* **2025**.
145. 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 CVPR, 2025.
146. 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 CVPR, 2025, pp. 17337–17346.
147. 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. In Proceedings of the CVPR, 2025.
148. Corsetti, J.; Giuliari, F.; Fasoli, A.; Boscaini, D.; Poiesi, F. Functionality understanding and segmentation in 3D scenes. In Proceedings of the CVPR, 2025, pp. 24550–24559.
149. Basak, H.; Yin, Z. ViSPLA: Visual Iterative Self-Prompting for Language-Guided 3D Affordance Learning. *NeurIPS* **2025**, *38*, 65706–65724.
150. Liu, S.C.; Chen, W.; Cheng, W.L.; Huang, Y.L.; Liao, I.B.; Li, Y.H.; Zhang, J.; et al. PAVLM: Advancing point cloud based affordance understanding via vision-language model. In Proceedings of the IROS, 2025, pp. 4299–4306.
151. Wu, L.; Wei, W.; Yu, P.; Lan, J. Open-Vocabulary 3D Affordance Understanding via Functional Text Enhancement and Multilevel Representation Alignment. In Proceedings of the ACM MM, 2025.
152. Li, D.; Feng, J.; Chen, J.; Dong, W.; Li, G.; Zheng, Y.; Feng, M.; Shi, G. SeqAffordSplat: Scene-level Sequential Affordance Reasoning on 3D Gaussian Splatting. *arXiv:2507.23772* **2025**.
153. 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. In Proceedings of the IROS, 2025, pp. 9433–9439.
154. Wei, Z.; Lin, J.; Liu, Y.; Chen, W.; Luo, J.; Li, G.; Lin, L. 3daffordsplat: Efficient affordance reasoning with 3d gaussians. In Proceedings of the ACM MM, 2025, pp. 2821–2830.
155. 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. In Proceedings of the AAAI, 2026.
156. He, L.; Liu, M.; Ye, Q.; Zhou, Y.; Deng, X.; Ding, G. Task-Aware 3D Affordance Segmentation via 2D Guidance and Geometric Refinement. In Proceedings of the AAAI, 2026.
157. Lee, S.I.; Lei, Q.; Xu, D.; Yang, D.; Tan, R.T.; Chen, Y.; Wang, B. Token-Based Affordance Grounding with Large Vision-Language Models. *arXiv:2607.03595* **2026**.
158. Morin, S.; Gupta, K.; Sandhu, M.; Gauthier, C.; Argenziano, F.; Ellis, K.; Paull, L. Agentic Scene Policies: Unifying Space, Semantics, and Affordances for Robot Action. *arXiv:2509.19571* **2025**.
159. Maillard, L.; Engelmann, F.; Durand, T.; Pan, B.; You, Y.; Litany, O.; Guibas, L.; Ovsjanikov, M. SceneTeract: Agentic Functional Affordances and VLM Grounding in 3D Scenes. *arXiv:2603.29798* **2026**.
160. Li, D.; Feng, J.; Li, G.; Shang, R.; Zheng, Y.; Dong, W.; Shi, G. A3R: Agentic Affordance Reasoning via Cross-Dimensional Evidence in 3D Gaussian Scenes. *arXiv:2604.01882* **2026**.
161. Huang, H.; Shi, J.; Li, Y.; Chen, Y. Affordance Agent Harness: Verification-Gated Skill Orchestration. *arXiv:2605.00663* **2026**.
162. Wang, Q.; He, J.; Pan, Y.; Yang, X.; Li, S. Grounding by Remembering: Cross-Scene and In-Scene Memory for 3D Functional Affordances. *arXiv:2605.11616* **2026**.
163. Zhang, Z.; Chen, K.; Wang, H.; Zhang, H.; Chen, H.H.; Liao, C.; Guo, L.; Chen, Y.C. A4-Agent: An Agentic Framework for Zero-Shot Affordance Reasoning. *arXiv:2512.14442* **2025**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

164. Mees, O.; Abdo, N.; Mazuran, M.; Burgard, W. Metric Learning for Generalizing Spatial Relations to New Objects. In Proceedings of the IROS, Vancouver, Canada, 2017.
165. Mees, O.; Burgard, W. Composing Pick-and-Place Tasks By Grounding Language. In Proceedings of the ISER, La Valletta, Malta, 2021.
166. Zhou, D.; He, M.; Fang, Z.; Yao, X.; Liu, Y.; Knoll, A.; Cao, H. AffordanceGrasp-R1: Leveraging Reasoning-Based Affordance Segmentation with Reinforcement Learning for Robotic Grasping. *arXiv:2602.03547* **2026**.
167. Wang, J.; Li, W.; Wu, Y.; Liang, Y.; Guo, Y.; Li, C.; Duan, H.; Zhang, Z.; Zhai, G. Affordance Benchmark for MLLMs. *arXiv:2506.00893* **2025**.
168. Wan, Z.; Xie, Y.; Zhang, C.; Lin, Z.; Wang, Z.; Stepputtis, S.; Ramanan, D.; Sycara, K. InstructPart: Task-Oriented Part Segmentation with Instruction Reasoning. In Proceedings of the ACL, 2025.
169. Wu, D.; Fu, Y.; Huang, S.; Liu, Y.; Jia, F.; Liu, N.; Dai, F.; Wang, T.; Anwer, R.M.; Khan, F.S.; et al. RAGNet: Large-scale Reasoning-based Affordance Segmentation Benchmark towards General Grasping. In Proceedings of the ICCV, 2025, pp. 11980–11990.
170. Tang, Y.; Zhang, L.; Zhang, S.; Zhao, Y.; Hao, X. Roboafford: A dataset and benchmark for enhancing object and spatial affordance learning in robot manipulation. In Proceedings of the ACM MM, 2025, pp. 12706–12713.
171. Hao, X.; Tang, Y.; Zhang, L.; Ma, Y.; Diao, Y.; Jia, Z.; Ding, W.; Ye, H.; Chen, L. RoboAfford++: A Generative AI-Enhanced Dataset for Multimodal Affordance Learning in Robotic Manipulation and Navigation. *arXiv:2511.12436* **2025**.
172. Gou, D.; He, X. Part-Aware Open-Vocabulary 3D Affordance Grounding via Prototypical Semantic and Geometric Alignment. *arXiv:2603.17647* **2026**.
173. Xu, R.; Shen, Y.; Li, X.; Wu, R.; Dong, H. NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation. *IEEE RA-L* **2024**.
174. Jia, W.; Li, K.; Yang, F.; Duan, M.; Chen, W.; Jiang, Y.; Zhang, H.; Yang, K.; Li, Z.; Wang, Y. TRACER: Texture-Robust Affordance Chain-of-Thought for Deformable-Object Refinement. *arXiv:2601.20208* **2026**.
175. Wu, H.; Zhang, Z.; Cheng, H.; Yang, K.; Liu, J.; Guo, Z. Learning affordance space in physical world for vision-based robotic object manipulation. In Proceedings of the ICRA, 2020, pp. 4652–4658.
176. Mo, K.; Guibas, L.J.; Mukadam, M.; Gupta, A.; Tulsiani, S. Where2act: From pixels to actions for articulated 3d objects. In Proceedings of the ICCV, 2021, pp. 6813–6823.
177. Wu, R.; Zhao, Y.; Mo, K.; Guo, Z.; Wang, Y.; Wu, T.; Fan, Q.; Chen, X.; Guibas, L.; Dong, H. VAT-Mart: Learning Visual Action Trajectory Proposals for Manipulating 3D ARTiculated Objects. In Proceedings of the ICLR, 2022.
178. Wang, Y.; Wu, R.; Mo, K.; Ke, J.; Fan, Q.; Guibas, L.J.; Dong, H. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In Proceedings of the ECCV, 2022, pp. 90–107.
179. Zhao, Y.; Wu, R.; Chen, Z.; Zhang, Y.; Fan, Q.; Mo, K.; Dong, H. DualAfford: Learning Collaborative Visual Affordance for Dual-gripper Manipulation. In Proceedings of the ICLR, 2023.
180. Wu, R.; Cheng, K.; Zhao, Y.; Ning, C.; Zhan, G.; Dong, H. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *NeurIPS* **2023**, *36*, 60966–60983.
181. Wu, R.; Ning, C.; Dong, H. Learning foresightful dense visual affordance for deformable object manipulation. In Proceedings of the ICCV, 2023, pp. 10947–10956.
182. Zhang, X.; Wang, D.; Han, S.; Li, W.; Zhao, B.; Wang, Z.; Duan, X.; Fang, C.; Li, X.; He, J. Affordance-Driven Next-Best-View Planning for Robotic Grasping. In Proceedings of the CoRL, 2023.
183. Di Palo, N.; Johns, E. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In Proceedings of the ICRA, 2024, pp. 2798–2805.
184. Ding, K.; Chen, B.; Wu, R.; Li, Y.; Zhang, Z.; Gao, H.a.; Li, S.; Zhou, G.; Zhu, Y.; Dong, H.; et al. Preafford: Universal affordance-based pre-grasping for diverse objects and environments. In Proceedings of the IROS, 2024, pp. 7278–7285.
185. Xu, R.; Zhang, J.; Guo, M.; Wen, Y.; Yang, H.; Lin, M.; Huang, J.; Li, Z.; Zhang, K.; Wang, L.; et al. A0: An affordance-aware hierarchical model for general robotic manipulation. *arXiv:2504.12636* **2025**.
186. Kim, T.; Bae, H.; Li, Z.; Li, X.; Ponomarenko, I.; Wu, R.; Dong, H. ManipGPT: Is Affordance Segmentation by Large Vision Models Enough for Articulated Object Manipulation? In Proceedings of the IROS, 2025, pp. 20974–20981.
187. Wei, Y.L.; Lin, M.; Lin, Y.; Jiang, J.J.; Wu, X.M.; Zeng, L.A.; Zheng, W.S. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. In Proceedings of the ICCV, 2025, pp. 11818–11828.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

188. Wang, S.; Yang, Y.; Luo, Y.; Li, D.; Wei, W.; Zhang, Y.; Hu, P.; Fu, Y.; Duan, H.; Sun, J.; et al. ScaleADFG: Affordance-based Dexterous Functional Grasping via Scalable Dataset. *IEEE RA-L* **2025**.
189. Shen, Y.; Wu, R.; Ke, Y.; Song, X.; Li, Z.; Li, X.; Fan, H.; Lu, H.; et al. BiAssemble: Learning Collaborative Affordance for Bimanual Geometric Assembly. *arXiv:2506.06221* **2025**.
190. Wu, R.; Zhu, Z.; Wang, Y.; Chen, Y.; Wang, J.; Dong, H. GarmentPile: Point-level visual affordance guided retrieval and adaptation for cluttered garments manipulation. In Proceedings of the CVPR, 2025, pp. 6950–6959.
191. Liang, J.; Chen, Y.; Yu, Q.; Shen, Y.; Zhang, H.; Dong, H.; Wu, R. A3D: Adaptive Affordance Assembly with Dual-Arm Manipulation. *arXiv:2601.11076* **2026**.
192. Shen, Y.; Jiang, F.; He, Z.; Li, X.; Liu, Y.; Li, Z.; Wu, R.; Dong, H. BiPreManip: Learning Affordance-Based Bimanual Preparatory Manipulation through Anticipatory Collaboration. In Proceedings of the CVPR, 2026, pp. 42430–42440.
193. Zhuang, Q.; Xu, H.Y.; Wang, Y.; Zhao, X.Y.; Li, Y.Y.; Wei, X.S. RAAP: Retrieval-Augmented Affordance Prediction with Cross-Image Action Alignment. *arXiv:2603.29419* **2026**.
194. Han, Y.; Peng, Y.; Yi, P.; Li, J.; Wang, H.; Zhang, G.; Liu, Q.P.; Lian, W. FSAG: Enhancing Human-to-Dexterous-Hand Finger-Specific Affordance Grounding via Diffusion Models. *arXiv:2601.08246* **2026**.
195. Xu, S.; Wang, Z.; Wang, Y.; Xia, C.; Huang, T.; Xu, C. Affordance Field Intervention: Enabling VLAs to Escape Memory Traps in Robotic Manipulation. In Proceedings of the CVPR, 2026, pp. 37206–37215.
196. Khazatsky, A.; Nair, A.; Jing, D.; Levine, S. What can i do here? learning new skills by imagining visual affordances. In Proceedings of the ICRA, 2021, pp. 14291–14297.
197. Mandikal, P.; Grauman, K. Learning dexterous grasping with object-centric visual affordances. In Proceedings of the ICRA, 2021, pp. 6169–6176.
198. Geng, Y.; An, B.; Geng, H.; Chen, Y.; Yang, Y.; Dong, H. Rlafford: End-to-end affordance learning for robotic manipulation. In Proceedings of the ICRA, 2023, pp. 5880–5886.
199. Bharadhwaj, H.; Gupta, A.; Tulsiani, S. Visual affordance prediction for guiding robot exploration. In Proceedings of the ICRA, 2023, pp. 3029–3036.
200. Zhang, H.; Kan, Z.; Shang, W.; Song, Y. A Novel Task-Driven Diffusion-Based Policy With Affordance Learning for Generalizable Manipulation of Articulated Objects. *IEEE/ASME Transactions on Mechatronics* **2025**.
201. Ma, T.; Zheng, J.; Wang, Z.; Gao, Z.; Zhou, J.; Liang, J. Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation. *arXiv:2505.11865* **2025**.
202. Zhang, L.; Mondal, S.; Bing, Z.; Bai, K.; Zheng, D.; Chen, Z.; Knoll, A.C.; Zhang, J. DORA: Object affordance-guided reinforcement learning for dexterous robotic manipulation. In Proceedings of the CBS, 2025, pp. 674–681.
203. Zhao, H.; Zhuang, L.; Zhao, X.; Zeng, C.; Xu, H.; Jiang, Y.; Cen, J.; Wang, K.; Guo, J.; Huang, S.; et al. Towards affordance-aware robotic dexterous grasping with human-like priors. In Proceedings of the AAAI, 2026, Vol. 40, pp. 13126–13134.
204. 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:2606.04172* **2026**.
205. Chen, W.; Liang, H.; Chen, Z.; Sun, F.; Zhang, J. Learning 6-dof task-oriented grasp detection via implicit estimation and visual affordance. In Proceedings of the IROS, 2022, pp. 762–769.
206. Ruan, S.; Liu, W.; Wang, X.; Meng, X.; Chirikjian, G.S. PRIMP: Probabilistically-informed motion primitives for efficient affordance learning from demonstration. *IEEE T-RO* **2024**, *40*, 2868–2887.
207. Han, Y.; Liu, J.; Zhang, H.; Gu, Y.; Guo, Y.; Lian, W. BridgeACT: Bridging Human Demonstrations to Robot Actions via Unified Tool-Target Affordances. *arXiv:2604.23249* **2026**.
208. van Oort, T.; Miller, D.; Browne, W.N.; Marticorena, N.; Haviland, J.; Suenderhauf, N. Open-vocabulary part-based grasping. *arXiv:2406.05951* **2024**.
209. Tong, E.; Balaji, A.; Opipari, A.; Lewis, S.; Zeng, Z.; Jenkins, O.C. OVAL-Grasp: Open-Vocabulary Affordance Localization for Task Oriented Grasping. *arXiv:2511.20841* **2025**.
210. Nagarajan, T.; Grauman, K. Learning affordance landscapes for interaction exploration in 3d environments. *NeurIPS* **2020**, *33*, 2005–2015.
211. Zhang, F.; Gienger, M. Affordance-based robot manipulation with flow matching. *arXiv:2409.01083* **2024**.
212. Cruz, F.; Magg, S.; Weber, C.; Wermter, S. Training agents with interactive reinforcement learning and contextual affordances. *IEEE TCDS* **2016**, *8*, 271–284.
213. Nakamoto, M.; Mees, O.; Kumar, A.; Levine, S. Steering Your Generalists: Improving Robotic Foundation Models via Value Guidance. *Conference on Robot Learning (CoRL)* **2024**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0392.v1

214. Guo, A.; Wen, B.; Yuan, J.; Tremblay, J.; Tyree, S.; Smith, J.; Birchfield, S. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In Proceedings of the IROS, 2023, pp. 11428–11435.
215. Zhang, C.; Delitzas, A.; Wang, F.; Zhang, R.; Ji, X.; Pollefeys, M.; Engelmann, F. Open-vocabulary functional 3d scene graphs for real-world indoor spaces. In Proceedings of the CVPR, 2025, pp. 19401–19413.
216. Heidinger, M.; Jauhri, S.; Prasad, V.; Chalvatzaki, G. 2handedafforder: Learning precise actionable bimanual affordances from human videos. In Proceedings of the ICCV, 2025, pp. 14743–14753.
217. Yu, Z.; Huang, Y.; Furuta, R.; Yagi, T.; Goutsu, Y.; Sato, Y. Fine-grained affordance annotation for egocentric hand-object interaction videos. In Proceedings of the WACV, 2023, pp. 2155–2163.
218. Li, M.; Xu, H.; Sun, H.; Chen, X.; Ren, S.; Huang, L.; Sui, X.; Miao, C.; Ye, J.; Cui, Q.; et al. AffordSim: A Scalable Data Generator and Benchmark for Affordance-Aware Robotic Manipulation. *arXiv:2604.11674* **2026**.
219. Blank, N.; Reuss, M.; Rühle, M.; Yağmurlu, Ö.E.; Wenzel, F.; Mees, O.; Lioutikov, R. Scaling Robot Policy Learning via Zero-Shot Labeling with Foundation Models. *Conference on Robot Learning (CoRL)* **2024**.
220. Zhang, J.; Hu, K.; Huang, Y.; Ju, Y.; Xue, Z.; Xu, H. AffordGen: Generating Diverse Demonstrations for Generalizable Object Manipulation with Affordance Correspondence. In Proceedings of the CVPR, 2026, pp. 15966–15975.
221. He, X.; Polavaram, A.; Cao, Y.; Deshmukh, O.; Wang, T.; Zhou, X.; Fang, K. Generate, Transfer, Adapt: Learning Functional Dexterous Grasping from a Single Human Demonstration. *arXiv:2601.05243* **2026**.
222. Hou, B.; Li, G.; Jia, J.; An, T.; Guo, X.; Leng, S.; Geng, H.; Ze, Y.; Harada, T.; Torr, P.; et al. World Model for Robot Learning: A Comprehensive Survey. *IJRR* **2026**.
223. Li, G.; Zhao, B.; Yang, J.; Sevilla-Lara, L. Mask2iv: Interaction-centric video generation via mask trajectories. In Proceedings of the AAAI, 2026, Vol. 40, pp. 6091–6099.
224. Deng, Y.; Liu, Y.; Hsu, D. GROW2: Grounding Which and Where for Robot Tool Use. *arXiv:2606.30632* **2026**.
225. Mees, O.; Hermann, L.; Rosete-Beas, E.; Burgard, W. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE RA-L* **2022**, *7*, 7327–7334.
226. Hu, Y.; Chen, X.; Zhou, C.; Liu, Y.; Li, G.; Jia, J.; Yang, J. Trace-Focused Diffusion Policy for Multi-Modal Action Disambiguation in Long-Horizon Robotic Manipulation. *arXiv:2602.07388* **2026**.
227. Mazzaglia, P.; Cohen, T.; Dijkman, D. Information-driven affordance discovery for efficient robotic manipulation. In Proceedings of the ICRA, 2024, pp. 7780–7787.
228. Nematollahi, I.; Mees, O.; Hermann, L.; Burgard, W. Hindsight for Foresight: Unsupervised Structured Dynamics Models from Physical Interaction. In Proceedings of the IROS, Las Vegas, USA, 2020.
229. Li, X.; Hsu, K.; Gu, J.; Pertsch, K.; Mees, O.; Walke, H.R.; Fu, C.; Lunawat, I.; Sieh, I.; Kirmani, S.; et al. Evaluating Real-World Robot Manipulation Policies in Simulation. In Proceedings of the CoRL, 2024.
230. Jones, J.; Mees, O.; Sferrazza, C.; Stachowicz, K.; Abbeel, P.; Levine, S. Beyond Sight: Finetuning Generalist Robot Policies with Heterogeneous Sensors via Language Grounding. In Proceedings of the ICRA, Atlanta, USA, 2025.
231. Shan, S.; Zhou, C.; Wang, R.; Chen, X.; Chen, X.; Zhou, X.; Ma, B.; Hu, I.Y.; Li, J.; Hu, C.Y.; et al. Learning Physical Interaction: A Survey of Tactile-and Force-aware Robot Learning. *arXiv:2608.07558* **2026**.
232. Mon-Williams, R.; Li, G.; Long, R.; Du, W.; Lucas, C.G. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence* **2025**, pp. 1–10.
233. Liu, P.; Li, G.; Fan, J.; Ma, B.; Jia, J.; Xiao, Y.; Yang, J. GIVE: Grounding Human Gestures in Vision-Language-Action Models. *arXiv:2606.13435* **2026**.
234. Zuo, K.; Li, G.; Lyu, B.; Lu, Y.; Ma, B.; Han, S.; Zhou, X.; Yuan, X.; Zhou, C.; Bai, J.; et al. Gaze2act: Gaze-conditioned vision-language-action policies for interactive robot manipulation. *arXiv:2605.30282* **2026**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.