

Affogato: Open-Vocabulary Affordance Grounding with Automated Data Generation at Scale

Junha Lee^{1,2*}, Eunha Park^{1*}, Chunghyun Park¹,
Dahyun Kang¹, and Minsu Cho^{1,3}

¹ Pohang University of Science and Engineering (POSTECH)

² SqueezeBits

³ RLWLRD

<https://junha-l.github.io/affogato/>

Abstract. Affordance grounding aims to localize where to interact with an object, a fundamental capability for embodied agents. Yet progress is bottlenecked by data: manual annotation is prohibitively expensive and confines existing datasets to a narrow set of predefined object and affordance categories. We introduce **Affogato**, a framework for open-vocabulary affordance grounding centered on **Affogato-750K**, a large-scale dataset of 750K 3D affordance heatmaps paired with natural language queries. We build it with a fully automated pipeline that orchestrates foundation models to generate them at scale without human labeling. It covers significantly more diverse categories than any existing dataset. For reliable evaluation, we further provide 5K human-verified test pairs. We also present **Espresso-3D** and **Espresso-2D**, simple yet effective models with a unified architecture across both modalities. Pretraining on Affogato-750K improves both Espresso and prior methods and yields the largest gains on unseen object and affordance categories, showing that it provides broadly transferable supervision across architectures.

Keywords: affordance grounding, dataset and benchmark · open-vocabulary · automated annotation

1 Introduction

The theory of affordances addresses what an environment offers, provides, or furnishes animals [18]. For example, what does a cup afford to humans? The answer can be drinking, lifting, throwing, *etc.* While the concept of affordances has long been studied in psychology [2, 25] and robotics [1, 5, 44], here we focus on *visual affordance grounding* from a computer vision perspective [20]; given an object as a 3D point cloud or 2D image, the task is to localize the region where a described interaction takes place. Unlike other types of visual grounding/segmentation [28, 31, 71], affordance grounding is particularly challenging for two main reasons. First, an object may afford multiple functionalities across different regions, and these regions often have indistinct boundaries, making localization inherently

*Equal contribution.

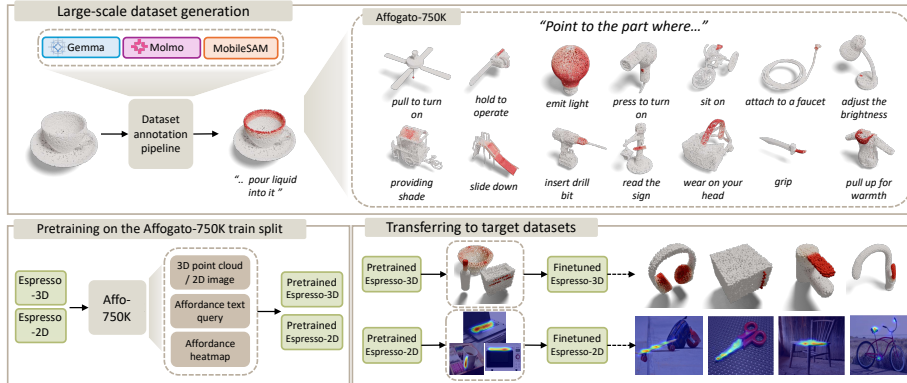


Fig. 1: Overview of Affogato. At the heart of Affogato is **Affogato-750K** (dataset), validated through **Espresso** (models). Affogato-750K comprises 750K open-vocabulary affordance annotations across 150K 3D objects, built via an automated pipeline chaining Gemma3 [17] (query generation), Molmo [12] (point prediction), and MobileSAM [72] (mask generation) with multi-view voting for 3D aggregation. The test split is human-verified to ensure evaluation quality. Espresso-3D and Espresso-2D follow a unified architecture and demonstrate substantial pretraining gains for Espresso models and existing methods.

ambiguous. Second, affordances are open-vocabulary by nature, and a fixed set of predefined terms cannot describe them adequately.

Addressing these challenges requires affordance supervision that is open-vocabulary, captures the soft and ambiguous extent of affordance regions, and can be generated at scale without exhaustive manual annotation. However, existing datasets fail to capture this open-ended nature; current 3D datasets [8, 15, 37, 51, 68, 70] are constrained to a small set of predefined object and affordance categories, and 2D datasets [9, 16, 41, 47–50] remain similarly limited in coverage (Table 1). This limitation stems from their reliance on manual annotation, which cannot scale to the diversity of real-world objects and interactions.

To overcome these limitations, we introduce **Affogato** (*AFFOrdance Grounding All aT Once*), a scalable framework for open-vocabulary affordance grounding (Figure 1). At its core is **Affogato-750K**, a large-scale dataset that directly addresses the scarcity of open-vocabulary affordance data. It is built fully automatically with an annotation pipeline that orchestrates the complementary strengths of multiple foundation models [12, 17, 72]. The pipeline generates open-vocabulary affordance queries, localizes and segments the corresponding interaction regions in each view, and aggregates these per-view predictions into robust 3D heatmaps.

Affogato-750K comprises 150K 3D objects from Objaverse [13], each annotated with five affordance queries and corresponding 3D heatmaps, totaling 750K annotations. We further derive a 2D split by projecting the heatmaps onto rendered views, and curate a test split of 5K human-verified pairs for reliable

Table 1: Comparison of 3D (left) and 2D (right) affordance grounding datasets. Aff abbreviates affordances. For Affogato-750K, classes and affordances are Gemma3 [17] predictions with frequencies above 10 and 100, respectively.

Dataset	# Classes	# Aff	# 3D assets	# Questions	Dataset	# Classes	# Aff	# 2D images	# Questions
3D AffordanceNet [15]	23	18	22,949	0	UMD [48]	17	7	10,000	0
PartAfford [63]	23	24	>25,000	0	IIT-Aff [50]	10	9	8,835	0
OpenAD [51]	23	37	22,949	0	ADE-Aff [9]	150	3	10,011	0
PIAD [68]	23	17	7,012	0	ORPA [16]	N/A	7	2,512	0
LASO [37]	23	17	8,434	870	Grounded I.H. [49]	31	20	1,871	0
SceneFun3D [14]	N/A	9	710	17,133	AGD20K [41]	50	36	23,816	0
IRAS [8]	23	36	22,949	42,119	EPIC-Aff [47]	304	20-36	38,876	0
SeqAfford [70]	23	18	18,371	162,386	3DOI [55]	N/A	3	10,000	0
Affogato-750K	>450	>350	150,104	750,520	VRB [3]	N/A	N/A	54,000	0
					2HandedAfforder [22]	163	73	278,000	278,000
					RAGNet [62]	180	N/A	273,000	26,000
					Affogato-750K	>450	>350	150,104	750,520

evaluation. As shown in Table 1, Affogato-750K significantly surpasses existing datasets in both scale and category diversity.

Alongside the dataset, we design Espresso-3D and Espresso-2D, minimal and symmetric models built from a vision encoder, a text encoder, and a heatmap decoder. Despite this simplicity, Espresso achieves strong affordance grounding performance in both modalities. Experiments demonstrate that pretraining on Affogato-750K yields substantial and consistent gains for both Espresso and existing methods, confirming the dataset’s effectiveness as a scalable supervision source across modalities.

To summarize, our contributions are as follows:

- We introduce **Affogato**, a scalable framework for open-vocabulary affordance grounding, built on **Affogato-750K**, a large-scale dataset of 750K affordance annotations over 150K 3D instances that covers far more diverse object and affordance categories than existing datasets, with a 5K human-verified test split for reliable evaluation.
- We develop a fully automated pipeline that orchestrates foundation models to generate reliable query-heatmap pairs at scale without manual labeling.
- We present **Espresso-3D** and **Espresso-2D**, simple yet effective models with a unified design across 3D and 2D, and show that pretraining on Affogato-750K yields substantial gains for both Espresso and prior methods.

2 Related work

Visual affordance grounding. Visual affordance grounding aims to localize functional regions in visual data given affordance concepts. Early 2D approaches [3, 9, 16, 41, 47–50, 55] focused on predefined affordance categories with limited object diversity. To overcome the scarcity of annotations, recent works explored weakly supervised learning [41, 59] and few-shot or zero-shot approaches [11, 34]. Other works leveraged external videos of human–object interactions [16, 22, 27, 40, 47, 49] to generate affordance heatmaps. However, supervision derived from such videos is often noisy and biased due to occlusions and limited viewpoints. Although recent advances in automation have enabled the construction of large-scale

2D datasets [22, 62], annotations based solely on single-view observations are inherently constrained compared to annotations aggregated from multi-view evidence.

In the 3D domain, 3D AffordanceNet [15] established the first closed-vocabulary 3D affordance grounding benchmark using PartNet [46] shapes. PartAfford [63] and OpenAD [51] expanded affordance vocabularies while maintaining categorical constraints. Recent advances toward open-vocabulary 3D grounding include LASO [37], which introduced free-form textual queries with 3D heatmap prediction, and 3D AffordanceLLM [8], which leveraged language models for query generation. SeqAfford [70] focused on textual diversity through GPT-4 generated questions, and LMAffordance3D [73] fused 2D and 3D features with language guidance. At the scene level, SceneFun3D [14] explored functionality segmentation, and Fun3DU [10] proposed a training-free framework leveraging VLMs for 3D scene affordance. More recently, methods that transfer 2D semantic knowledge to 3D affordance segmentation have also been explored [21, 23].

Despite progress, existing datasets remain severely constrained in scale and category diversity (Table 1). 3D datasets share the same 23 object classes derived from a single source [46], while 2D datasets cover at most 304 object and 73 affordance categories. All existing datasets rely on predefined closed-set labels, fundamentally limiting generalization to open-vocabulary queries. This data scarcity is the primary bottleneck for affordance grounding generalization.

Harnessing 2D foundation models for 3D supervision. Recent advances in 2D foundation models have opened new pathways for generating high-quality 3D supervision without manual annotation. Several works [43, 65, 66] leverage large language models (LLMs) and vision-language models (VLMs) to generate text captions for 3D objects using their multi-view rendered images, facilitating joint 3D-language learning. Others [39, 69] distill 2D visual knowledge from foundation models to train 3D geometric encoders, while scene-centric methods [32, 60, 67] extend this paradigm to large-scale environments by generating region-level 3D annotations through integration with 2D foundation models. Our work extends this paradigm to functional affordance understanding by leveraging 2D foundation models to generate 3D affordance supervision. A vision-language model generates affordance queries and interaction points, MobileSAM [72] produces 2D segmentation masks, and the resulting masks are lifted to 3D and aggregated via multi-view voting to obtain reliable 3D affordance heatmaps.

3 The Affogato-750K dataset

Advancing human-object interaction understanding in embodied AI systems demands comprehensive affordance grounding data—a critical resource currently lacking in the field. Existing datasets [15, 37] suffer from key limitations in scale, diversity, and annotation quality, creating a substantial barrier to progress in this domain. To address the issue, we propose the *Affogato-750K* dataset, a large-scale open-vocabulary affordance grounding dataset. In this section, we elaborate on the pipeline used to generate Affogato-750K. Our pipeline leverages large-scale

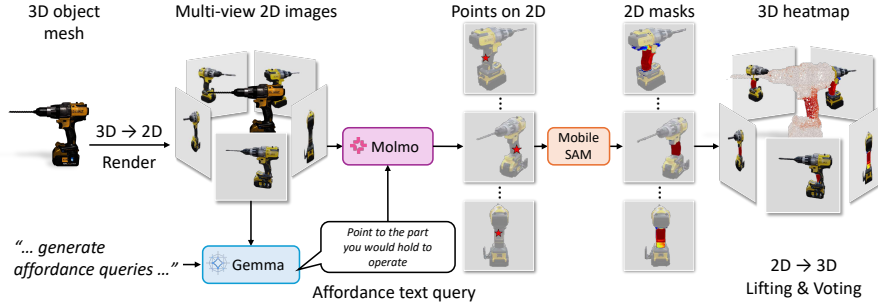


Fig. 2: Overview of our data annotation pipeline. Given multi-view renderings of an object, Gemma3 [17] generates affordance queries, Molmo [12] points the affordance, and MobileSAM [72] converts the point into an affordance heatmap. The multi-view 2D affordance heatmaps are then aggregated on the 3D object surface to obtain a final 3D affordance heatmap.

3D object repositories [13] and state-of-the-art foundation models [12, 17, 72] to automatically generate high-quality affordance annotations. This addresses the aforementioned limitations by enabling robust and generalizable affordance learning across diverse objects.

3.1 Source dataset

We build the Affogato-750K dataset upon Objaverse [13], one of the largest public 3D asset repositories. It contains more than 700K web-crawled 3D object meshes spanning diverse object categories and geometries. Since recent vision-language models (VLMs) take 2D images and text as input and cannot directly process 3D meshes, we use multi-view renderings of each 3D object as an intermediary representation that bridges 3D geometry and language. To this end, we utilize G-Objaverse [74], which already provides high-resolution multi-view renderings, along with depth maps and camera parameters, for over 280K Objaverse objects. Specifically, we select four subsets that have strong relevance to human-object interaction and functional affordances in daily lives: *Daily-Used*, *Furnitures*, *Transportations*, and *Electronics*, which results in 150K 3D objects.

3.2 Annotation pipeline

Our data annotation pipeline consists of three stages. Given a textured 3D object as input, it outputs a set of natural language affordance queries alongside spatially localized 3D affordance heatmaps. The entire process is automated and designed to scale to hundreds of thousands of objects, making it suitable for constructing large-scale datasets. The overall annotation process is illustrated in Figure 2. Stage-wise qualitative results are provided in the supplementary material.

Stage 1: Open-vocabulary affordance query generation. Given multi-view images of a 3D object, we employ a vision-language model, Gemma3 [17], to

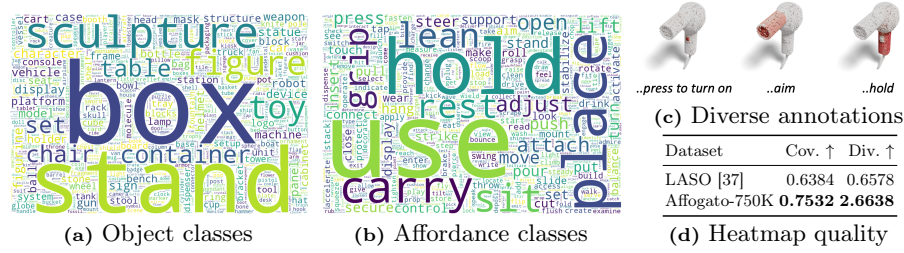


Fig. 3: Characteristics of the Affogato-750K dataset. (Left) The Affogato-750K exhibit high diversity in both object and affordance categories. (Right) Quantitative comparison of heatmap coverage and diversity with LASO, highlighting that Affogato-750K achieves significantly higher spatial coverage and distributional diversity (see Section 3.3 for details).

produce a set of natural language queries (five per object) that describe how a human might interact with the object. These queries follow a constrained yet expressive format, allowing open-vocabulary interaction descriptions while maintaining spatial grounding. By conditioning on rendered views instead of object class labels [8, 37, 70], our approach leverages the rich knowledge embedded in VLMs to generalize to open-vocabulary understanding of affordances. This also enables the system to adapt to various intra-class variations (*e.g.*, chairs with and without armrests) while maintaining consistent affordance identification across diverse object geometries and functional categories.

Stage 2: Language-guided interaction point prediction. Once the affordance queries are generated, we utilize Molmo [12], a multimodal model capable of grounding natural language queries to spatial locations in images. Specifically, Molmo demonstrates remarkable precision in identifying exact pixel locations when provided with input images and natural language prompts that request localization of specific regions in the image. This capability is crucial for accurately mapping affordance queries to their corresponding spatial locations on the object. For each query and image pair, we instruct Molmo to predict pixel coordinates that represent the most likely interaction point in the given view. The predicted interaction points across views are then used to guide the next stage.

Stage 3: Affordance heatmap generation and aggregation. We convert discrete interaction points into continuous 3D heatmap representations. First, the predicted point from each multi-view image is used as a prompt for MobileSAM [72] to generate a 2D segmentation mask. To prioritize the precise interaction region over broader contextual areas, we adopt a smallest-mask selection strategy. The segmentation logits are transformed through a sigmoid function to create probabilistic heatmaps with values between 0 and 1, representing the likelihood of an affordance at each pixel location. These 2D heatmaps from multiple viewpoints are then projected onto the 3D object surface using the known camera parameters and depth information. We employ a voting-based aggregation process where each view contributes to the final 3D representation, with regions consistently



Fig. 4: (Left) Human validation results from Section 3.3. (Right) Common failure cases in Affogato-750K: ● incorrect affordance queries from object misidentification; ● partial coverage of the target affordance region; ● SAM edge bias merging functionally distinct parts (e.g., backrest and seat); ● failure in fine-grained grounding of the target object.

identified across multiple views receiving higher confidence scores. The output of this stage serves as our final 3D affordance heatmap annotations.

Rendering 2D affordance heatmaps. To bridge our 3D dataset with 2D image domains, we project the aggregated 3D affordance heatmaps onto 2D image planes from multiple viewpoints. Since the multi-view aggregation in Stage 3 resolves the inconsistency of Molmo and MobileSAM predictions, these rendered 2D heatmaps exhibit high consistency across viewpoints with accurate affordance regions. For each object, we render heatmaps from 25 evenly distributed viewpoints and calculate affordance region visibility by summing the projected heatmap values. We select the most visible viewpoint for optimal representation.

Error mitigation. During the automatic dataset generation pipeline, we employ several strategies to mitigate errors. First, to enhance the fidelity of Gemma3’s responses, we adopt chain-of-thought (CoT) prompting [61], instructing the model to first predict the object’s semantic class and then generate affordance queries conditioned on that class and its functionalities. This approach prevents the model from relying solely on object shape, which often leads to affordance queries that fail to capture the object’s intended functions. Second, we apply a multi-view aggregation described in Stage 3 to reduce errors. Despite Molmo’s exceptional pointing capabilities, its predicted interaction points can still contain errors and vary across viewpoints. Although a few views produce incorrect heatmaps, the consensus aggregation reinforces the majority of correct predictions, resulting in robust, view-consistent outputs.

3.3 Statistics and analysis

Data statistics. As summarized in Table 1, Affogato-750K provides over 750K affordance annotations across 150K 3D object instances in four categories: *Daily-Used* includes the largest number of instances (121,799), followed by *Transportations* (11,609), *Furnitures* (8,759), and *Electronics* (7,937), each with 5 affordance query-heatmap pairs.

Semantic and spatial diversity. Unlike existing datasets constrained by pre-defined taxonomies, Affogato-750K spans a wide range of objects and affordances by leveraging VLM knowledge to generate diverse and context-aware affordance queries. Figures 3a and 3b demonstrate this semantic breadth across object classes and affordance types. Our annotations capture varied interaction patterns, from precise point interactions (*e.g.* button pressing) to extended surface interactions (*e.g.* holding), as shown in Figure 3c. Quantitatively, Table 3d shows Affogato-750K achieves significantly higher coverage (ratio of points covered by annotation union) and diversity (average pairwise KL divergence between heatmaps) scores than LASO [37], indicating richer and more complementary affordance representations.

Human validation and test split construction. To construct a reliable test split, we randomly sample 5K affordance query-heatmap pairs and assign them in disjoint subsets to 10 hired annotators for verification. We instruct annotators to rate each pair on semantic relevance, spatial accuracy, and heatmap coverage, and to inspect all stage-wise intermediates to locate the originating stage of any error. This evaluation shows strong agreement with human judgment, achieving an 84.8% success rate, and reveals the common failure modes of our pipeline (Figure 4). We then build the test split from these verified pairs: incorrect queries are removed and incorrect interaction points are re-annotated through a dedicated annotation interface, yielding an object-instance-level test pool that is held out from training with no instance, view, or query overlap.

4 Espresso models

Our goal is to isolate the contribution of Affogato-750K, attributing improvements in affordance grounding to the dataset itself rather than to model design. However, prior 2D [34, 41] and 3D [8, 15, 37] affordance methods have evolved in separate research streams, each accumulating its own modality-specific design choices. Adopting such methods as cross-modal baselines would conflate architectural differences with dataset contribution, making it difficult to cleanly attribute performance gains to Affogato-750K. We therefore present Espresso, a unified baseline that is symmetric across 2D and 3D, isolating the dataset’s contribution from architectural differences.

Concretely, we present Espresso-3D for 3D point clouds and Espresso-2D for 2D images. Each model consists of a modality-specific visual encoder, a text encoder, and a text-conditioned heatmap decoder (Figure 5). The core of our design is a text-conditioned heatmap decoder that uses text embeddings as its

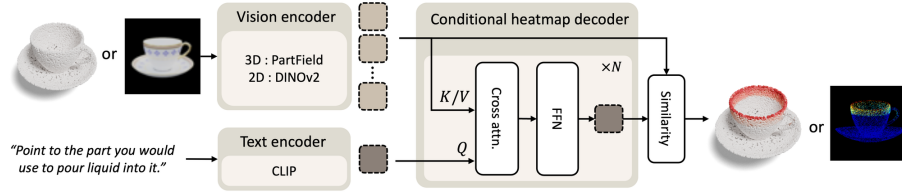


Fig. 5: Espresso-3D and Espresso-2D architectures. Each model consists of a 3D or 2D visual encoder, a text encoder, and a text-conditioned heatmap decoder. The affordance heatmap is predicted as the cross-modal similarity of the vision representation and the conditioned text embedding.

queries instead of learnable ones. Although this follows the standard transformer-based mask decoder architecture [6, 7, 24, 30, 58], using text embeddings as queries naturally supports open-vocabulary affordance grounding without predefined categories. The following describes the modality-specific designs.

Espresso-3D. For the 3D vision encoder, we use the pretrained PartField model [39], which captures generic part concepts from large-scale 3D data. For the text encoder, we employ Recap-CLIP [36], which provides robust language understanding capabilities. Following PointRefer [37], we finetune both the pretrained 3D vision encoder and text encoder during training to adapt them to our downstream task while leveraging their strong pretrained representations.

Espresso-2D. For the 2D vision encoder, we use the pretrained DINOv2 model [52], and CLIP [56] as the text encoder. Following OOAL [34], we freeze both encoders during training to maintain their pretrained representations. Our architecture adapts the design of OOAL [34] but differs in that we remove the learnable prompt component and take a single affordance query as input, following the referring expression segmentation framework.

5 Experiments

5.1 Datasets and baselines

3D datasets. We primarily evaluate Espresso-3D on the LASO dataset [37], a benchmark for open-vocabulary 3D affordance grounding with free-form text queries. LASO provides both seen and unseen settings; the seen setting evaluates on object/affordance classes observed during training, while the unseen setting tests on held-out classes to assess zero-shot generalization. We further evaluate on our Affogato-750K test split, a held-out set of 5K human-verified pairs that covers substantially more diverse object categories and affordance concepts than LASO. It serves as a more challenging testbed and provides both in-domain and cross-domain settings to probe open-vocabulary generalization beyond the limited category coverage of existing benchmarks.

2D datasets. We evaluate Espresso-2D on the AGD20K [41], which consists of two object splits: seen and unseen. AGD20K-Weak, 1-shot, and Full represent

Table 2: Open-vocabulary 3D affordance grounding on the LASO [37] test split. “Affogato-750K pretrain” denotes pretraining on the Affogato-750K train split and then finetuned on the LASO train split. Subscript values indicate the effect of Affogato-750K pretraining (green: improvement, gray: degradation).

Method	Seen				Unseen			
	aIoU↑	AUC↑	SIM↑	MAE↓	aIoU↑	AUC↑	SIM↑	MAE↓
Ref. Trans. [35]	13.7	79.8	49.7	0.124	10.2	69.1	43.2	0.145
3D-SPS [43]	11.4	76.2	43.3	0.138	7.9	68.8	40.2	0.158
ReLA [38]	15.2	78.9	53.2	0.118	10.7	69.7	42.9	0.144
IAGNet [68]	17.8	82.3	56.1	0.109	12.9	77.8	44.3	0.129
OpenAD [51]	14.2	85.1	53.3	0.103	14.6	80.7	51.8	0.109
+ Affogato-750K pretrain	16.1 _{+1.9}	86.8 _{+1.7}	53.9 _{+0.6}	0.100 _{-.003}	15.5 _{+0.9}	81.8 _{+1.1}	53.4 _{+1.6}	0.103 _{-.006}
PointRefer [37]	20.8	87.3	62.9	0.093	14.6	80.2	50.7	0.119
+ Affogato-750K pretrain	20.2 _{-0.6}	86.0 _{-1.3}	60.0 _{-2.9}	0.098 _{+0.005}	18.6 _{+4.0}	81.4 _{+1.2}	56.1 _{+5.4}	0.103 _{-.016}
Espresso-3D	20.4	86.0	63.3	0.102	18.7	80.0	60.0	0.101
+ Affogato-750K pretrain	21.9 _{+1.5}	85.9 _{-0.1}	63.7 _{+0.4}	0.116 _{+0.014}	20.8 _{+2.1}	82.9 _{+2.9}	61.4 _{+1.4}	0.122 _{+0.021}

weakly supervised, one-shot, and fully supervised versions of the AGD20K training dataset, respectively.

3D baselines. We select representative baselines for open-vocabulary 3D affordance grounding, which take language descriptions paired with 3D point clouds as input. Our evaluation includes multiple representative methods: Referring Transformer [35], 3D-SPS [43], ReLA [38], IAGNet [68], OpenAD [51], and PointRefer [37]. For OpenAD, since official results on LASO [37] were not available, we trained the model on the LASO dataset using their official code implementation. All other baseline results are taken directly from [37]. We compare these methods with our proposed Espresso-3D on both LASO seen and unseen splits to evaluate generalization performance.

2D baselines. For zero-shot evaluation on AGD20K, we use Molmo+SAM2 [12, 57], LISA-7B [31] (a reasoning segmentation model), M²SA-7B [26] (a part-level referring segmentation model), and AffordanceNet [62] (a reasoning-based affordance segmentation model). We format the open-vocabulary query for an affordance as “*Point to the part that you should interact with to {affordance}*”. For the affordance-specific model baselines in Table 4 (b), we adopt results from [54] and additionally include AffordanceNet [62].

Evaluation metrics. For 3D evaluation, following prior work [37], we use average Intersection over Union (aIoU), Area Under the ROC Curve (AUC), Similarity (SIM), and Mean Absolute Error (MAE). Since MAE is sensitive to annotation scale, we primarily report aIoU, AUC, and SIM. For 2D evaluation, we use Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) [4]. Higher values are better for aIoU, AUC, SIM, and NSS while lower values are better for MAE and KLD.

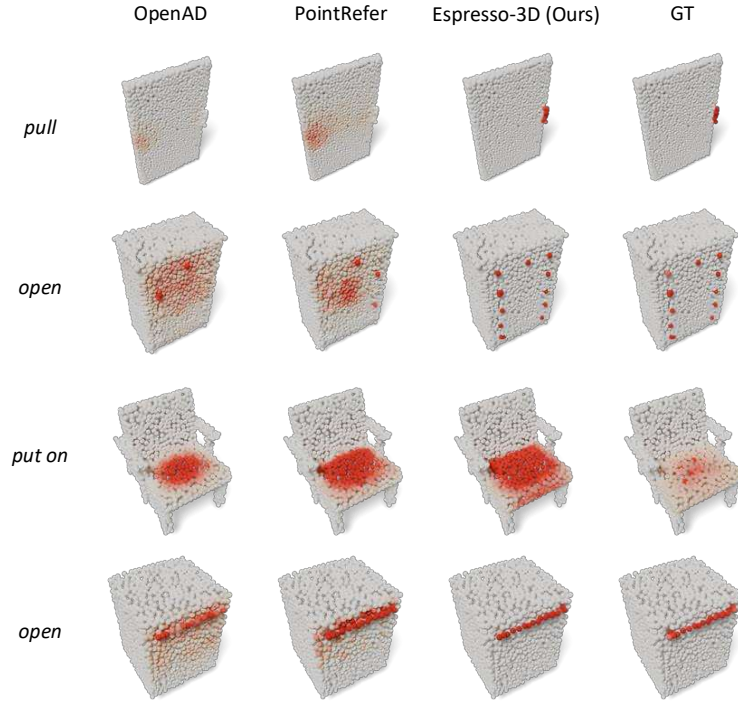


Fig. 6: Qualitative comparison between OpenAD [51], PointRefer [37], and Espresso-3D on the LASO [37] test split.

5.2 3D affordance grounding

LASO [37] test split. Table 2 presents the 3D affordance grounding results on LASO. Without pretraining on Affogato-750K, all methods achieve reasonable performance on the seen setting but show limited generalization to unseen object-affordance combinations. Pretraining on Affogato-750K consistently and substantially improves unseen performance across all methods: OpenAD gains 0.9%p aIoU, PointRefer achieves the largest gain of 4.0%p aIoU (18.6 vs. 14.6), and Espresso-3D gains 2.1%p aIoU. While pretrained PointRefer shows a slight degradation on the seen setting (20.2 vs. 20.8 aIoU), it achieves the largest unseen improvement; Espresso-3D improves consistently on both seen (+1.5%p aIoU) and unseen (+2.1%p aIoU). These results confirm that Affogato-750K pretraining provides a strong and consistent benefit for open-vocabulary generalization, regardless of model architecture. Figure 6 provides a qualitative comparison on the LASO test split.

Affogato-750K test split. We evaluate open-vocabulary 3D affordance grounding on the Affogato-750K test split under both in-domain and cross-domain settings (Table 3), comparing Espresso-3D against OpenAD [51] and PointRefer [37] retrained on our dataset. Since benchmarks like LASO cover few categories,

Table 3: Open-vocabulary 3D affordance grounding on the Affogato-750K test split. $All \rightarrow All$ denotes in-domain evaluation, while $Daily-used \rightarrow Furnitures$ and $Furnitures \rightarrow Daily-used$ represent cross-domain settings.

Method	$All \rightarrow All$				$Daily-used \rightarrow Furnitures$				$Furnitures \rightarrow Daily-used$			
	aIoU $\uparrow$	AUC $\uparrow$	SIM $\uparrow$	MAE $\downarrow$	aIoU $\uparrow$	AUC $\uparrow$	SIM $\uparrow$	MAE $\downarrow$	aIoU $\uparrow$	AUC $\uparrow$	SIM $\uparrow$	MAE $\downarrow$
OpenAD	3.1	64.8	32.9	0.150	7.6	67.8	36.8	0.177	1.6	54.3	30.8	0.201
PointRefer	10.5	76.1	40.5	0.120	13.5	78.9	44.3	0.157	2.8	60.6	26.5	0.122
Espresso-3D	13.6	79.0	42.9	0.111	18.2	80.4	47.5	0.125	4.6	62.4	30.4	0.130

this experiment tests generalization under the broader and more diverse distribution of Affogato-750K. The in-domain split ($All \rightarrow All$) trains and evaluates on all objects, while the cross-domain splits ($Daily-used \rightarrow Furnitures$ and $Furnitures \rightarrow Daily-used$) train on one object domain and test on the other to measure cross-domain transfer. Even the strongest methods leave substantial headroom on Affogato-750K, reflecting the broad object and affordance coverage over which it requires generalization. Across all three settings, Espresso-3D consistently outperforms OpenAD and PointRefer, mirroring their ranking on LASO and indicating that these trends are not benchmark-specific. The cross-domain results are strongly asymmetric: training on diverse daily-used objects and testing on furniture ($Daily-used \rightarrow Furnitures$) far outperforms the reverse (18.2 vs. 4.6 aIoU for Espresso-3D), since the narrow furniture domain covers little of the broader daily-used categories.

5.3 2D affordance grounding

Table 4 (a) presents results on the test sets of AGD20K’s seen and unseen object splits, evaluated zero-shot without any AGD20K training. Each baseline is pretrained on its own respective dataset, while Espresso-2D is pretrained on Affogato-750K. Espresso-2D with Affogato-750K pretraining outperforms all baselines across both test sets, achieving 40.2 SIM on the seen split and 37.6 SIM on the unseen split versus 22 to 30 SIM for competing methods, demonstrating that Affogato-750K pretraining yields affordance representations that generalize well across diverse object categories. Table 4 (b) presents supervised learning results across different supervision levels. Espresso-2D with Affogato-750K pretraining, fine-tuned on AGD20K, achieves the best performance across all supervision settings. Notably, Affogato-750K pretraining provides consistent improvements over training solely on AGD20K—the Affogato-750K-pretrained Espresso-2D outperforms its from-scratch counterpart by 0.060 KLD and 1.6 SIM under full supervision, and AffordanceNet similarly benefits from Affogato-750K pretraining—validating that Affogato-750K’s diverse affordance concepts transfer across domain boundaries even when target-domain data is available. Figure 9 shows qualitative examples.

Table 4: Open-vocabulary 2D affordance grounding on the AGD20K [41] test split. “Affogato-750K pretrain” denotes pretraining on the Affogato-750K train split.

(a) Zero-shot results				(b) Supervised learning results			
Method	KLD↓	SIM↑	NSS↑	Method	Sup.	KLD↓	SIM↑ NSS↑
<i>Seen split</i>				Cross-view-AG [41]	Weak	1.787	28.5 0.829
Molmo+SAM2 [12, 57]	1.804	26.1	0.729	Cross-view-AG+ [42]		1.765	27.9 0.882
LISA-7B [31]	1.627	29.6	0.819	AffCorrs [19]		1.618	34.8 1.021
M ² SA-7B [26]	1.772	25.8	0.620	LOCATE [33]		1.405	37.2 1.157
AffordanceNet [62]	1.729	26.6	0.755	WSAG-PLSP [64]		1.153	43.7 1.418
Espresso-2D	1.426	40.2	0.985	OOAL [34]	1-shot	1.070	46.1 1.503
<i>Unseen split</i>				LOCATE-Sup [33]	Full	1.907	23.6 0.641
Molmo+SAM2 [12, 57]	1.953	22.6	0.718	LOCATE-Sup-OWL [33, 45]		1.927	23.4 0.624
LISA-7B [31]	1.830	25.6	0.765	AffordanceLLM [54]		1.463	37.7 1.070
M ² SA-7B [26]	1.925	22.7	0.657	AffordanceNet [62]		1.871	30.5 0.750
AffordanceNet [62]	1.932	22.6	0.700	+ Affogato-750K pretrain		1.721	33.7 0.852
Espresso-2D	1.571	37.6	1.016	Espresso-2D		1.034	50.3 1.550
				+ Affogato-750K pretrain		0.974	51.9 1.645

Table 5: Training-evaluation overlap analysis on LASO unseen and AGD20K unseen.

(a) LASO unseen split (3D)						(b) AGD20K unseen split (2D)			
Pretraining	Filtering	aIoU↑	AUC↑	SIM↑	MAE↓	Pretraining	Filtering	KLD↓	SIM↑ NSS↑
		18.7	80.0	60.0	0.101			1.034	50.3 1.550
✓		20.8	82.9	61.4	0.122	✓		0.974	51.9 1.645
✓	✓	20.2	82.4	61.1	0.119	✓	✓	1.016	52.0 1.631

5.4 Analysis

Effect of held-out category filtering. To check that the pretraining gains are not simply due to Affogato-750K covering categories that each benchmark holds out, we remove the pretraining samples that fall into those held-out classes and re-train. We query Gemma3 [17] to judge whether each sample’s class label falls into a benchmark’s held-out categories, and conservatively remove any potential match. For LASO, which holds out (object, affordance) pairs, we drop samples whose predicted pair appears in the evaluation set (5.46%); for AGD20K, which holds out object categories, we exclude samples assigned to those categories (14.0%). Since these predictions are imperfect, the filter likely over-removes, making the numbers a lower bound. Even so, filtered pretraining consistently surpasses training from scratch on both benchmarks (Table 5; on LASO, 20.2 vs. 18.7 aIoU, within 0.6 of full pretraining), confirming that Affogato-750K’s benefits come from diverse supervision at scale rather than from covering the held-out categories.

Synthetic-to-real generalization. Although Affogato-750K is built from clean, isolated Objaverse renderings, the pretrained Espresso-2D transfers to realistic, in-the-wild inputs. To probe a large domain gap, we evaluate Espresso-2D zero-shot on cluttered robot workspace images from Open X-Embodiment [53]. As shown in Figure 7, despite never observing such scenes during training, Espresso-2D



Fig. 7: Zero-shot Espresso-2D on real-world robot scenes. Despite the visual domain gap from Objaverse renderings, Espresso-2D pretrained on Affogato-750K localizes open-vocabulary affordances correctly across diverse, cluttered Open X-Embodiment [53] robot workspace images.

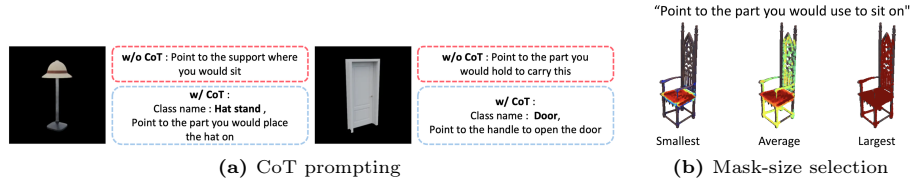


Fig. 8: Analyses of our data annotation pipeline. (a) CoT prompting produces functionally grounded affordance queries rather than shape-only guesses. (b) Among the three mask hypotheses returned by SAM, the smallest isolates part-level structure at affordance granularity.

still localizes the queried affordances correctly across diverse robot-manipulation scenes, indicating that the affordance representations learned from Affogato-750K generalize beyond the synthetic source domain.

Effect of chain-of-thought prompting. We qualitatively analyze the effect of CoT [61] prompting in our data annotation pipeline for affordance query generation. As shown in Figure 8a, without CoT prompting the model often fails to capture the object’s functional properties and instead relies primarily on shape when generating queries. For example, it produces “Point to the support where you would sit.” for a hat stand or “Point to the part you would hold to carry this.” for a door. In contrast, CoT prompting encourages functional reasoning, producing more appropriate queries such as “Point to the handle to open the door,” or “Point to the part you would place the hat on.”

Effect of mask-size selection. Another key design choice in our pipeline concerns how interaction points are turned into region masks. SAM [57, 72] returns three mask hypotheses for each interaction-point prompt, and we adopt the *smallest* one, as stated in Section 3.2. Figure 8b compares the three choices: the larger masks tend to snap to whole-object silhouettes, whereas the smallest mask isolates part-level structures (*e.g.*, handle, lid, button) at affordance granularity. Although the smallest mask can occasionally miss parts in individual views, our multi-view aggregation closes such per-view gaps, making the smallest-mask choice the most reliable basis for precise affordance heatmaps.

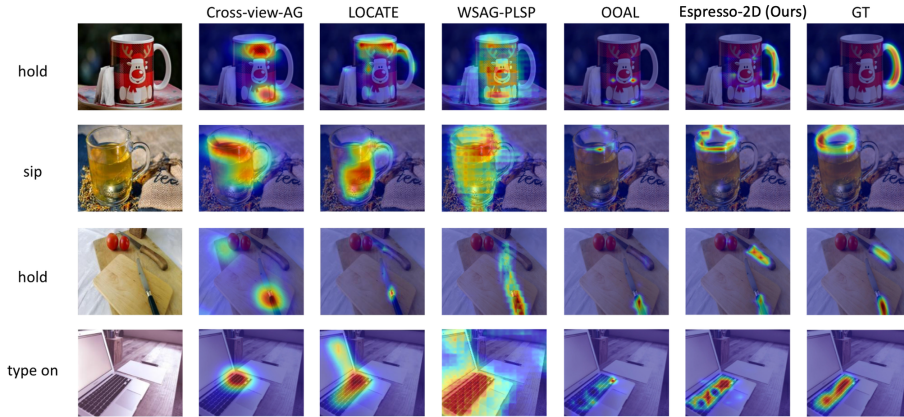


Fig. 9: Qualitative comparison between Cross-view-AG [41], LOCATE [33], WSAG-PLSP [64], OOAL [34], and Espresso-2D, which is pretrained on the Affogato-750K train split and then finetuned on the AGD20K train split with full supervision.

6 Discussion and conclusion

We have presented Affogato, a unified framework for open-vocabulary affordance grounding across 3D and 2D domains. By chaining foundation models, our automated pipeline produces Affogato-750K, the largest affordance grounding dataset with 750K affordance annotations across 150K 3D assets, eliminating costly manual annotation while capturing diverse interactions beyond predefined categories. Our Espresso-3D and Espresso-2D models consistently benefit from pretraining on Affogato-750K, demonstrating improved generalization to unseen object-affordance combinations. Notably, our annotation pipeline can be readily applied to other 3D datasets, offering a practical path toward ever-growing affordance supervision. While our framework effectively localizes contact points for a given action, predicting finer-grained information such as temporal dynamics and force requirements [29] remains an open challenge. We believe Affogato and Affogato-750K provide a solid foundation for future research toward more comprehensive embodied AI systems.

Acknowledgements

This work was supported by RLWORLD Inc. and also by IITP grants (RS-2022-II220959: Few-Shot Learning of Causal Inference in Vision and Language for Decision Making (50%), RS-2022-II220290: Visual Intelligence for Space-Time Understanding & Generation (25%), RS-2024-00457882: National AI Research Lab Project (20%), RS-2019-II191906: AI Graduate School Program at POSTECH (5%)) funded by Ministry of Science and ICT, Korea.

References

1. Ardón, P., Pairet, È., Lohan, K.S., Ramamoorthy, S., Petrick, R.: Affordances in robotic tasks—a survey. *arXiv preprint arXiv:2004.07400* (2020)
2. Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., Yoshida, C.: Cognitive developmental robotics: A survey. *IEEE transactions on autonomous mental development* **1**(1), 12–34 (2009)
3. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics. In: *CVPR*. pp. 13778–13790 (2023)
4. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? *IEEE TPAMI* **41**(3), 740–757 (2019)
5. Chemero, A., Turvey, M.T.: Gibsonian affordances for roboticists. *Adaptive Behavior* **15**(4), 473–480 (2007)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *CVPR*. pp. 1280–1289 (2022)
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. *NeurIPS* **34**, 17864–17875 (2021)
8. Chu, H., Deng, X., Lv, Q., Chen, X., Li, Y., Hao, J., Nie, L.: 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. In: *ICLR* (2025)
9. Chuang, C.Y., Li, J., Torralba, A., Fidler, S.: Learning to act properly: Predicting and explaining affordances from images. In: *CVPR*. pp. 975–983 (2018)
10. Corsetti, J., Giuliari, F., Fasoli, A., Boscaini, D., Poesi, F.: Functionality understanding and segmentation in 3d scenes. In: *CVPR*. pp. 24550–24559 (2025)
11. Cuttano, C., Rosi, G., Trivigno, G., Averta, G.: What does clip know about peeling a banana? In: *CVPRW*. pp. 2238–2247 (2024)
12. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: *CVPR*. pp. 91–104 (2025)
13. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., Vanderbilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: *CVPR*. pp. 13142–13153 (2023)
14. Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., Engelmann, F.: Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In: *CVPR*. pp. 14531–14542 (2024)
15. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: *CVPR*. pp. 1778–1787 (2021)
16. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: *CVPR*. pp. 2139–2147 (2018)
17. Gemma Team: Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025)
18. Gibson, J.J.: The ecological approach to visual perception: classic edition. Psychology press (2014)
19. Hadjivelichkov, D., Zwane, S., Agapito, L., Deisenroth, M.P., Kanoulas, D.: One-shot transfer of affordance regions? affcorrs! In: *CoRL*. pp. 550–560. PMLR (2023)
20. Hassanin, M., Khan, S., Tahtali, M.: Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)* **54**(3), 1–35 (2021)
21. He, L., Liu, M., Ye, Q., Zhou, Y., Deng, X., Ding, G.: Task-aware 3d affordance segmentation via 2d guidance and geometric refinement. In: *AAAI*. pp. 4654–4662 (2026)

22. Heidinger, M., Jauhri, S., Prasad, V., Chalvatzaki, G.: 2handedafforder: Learning precise actionable bimanual affordances from human videos. In: ICCV. pp. 14743–14753 (2025)
23. Huang, Y., Peng, Z., Wen, C., Yang, X., Shen, W.: Unlocking 3d affordance segmentation with 2d semantic knowledge. In: CVPR. pp. 6399–6409 (2026)
24. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: CVPR. pp. 2989–2998 (2023)
25. Jamone, L., Ugur, E., Cangelosi, A., Fadiga, L., Bernardino, A., Piater, J., Santos-Victor, J.: Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems* **10**(1), 4–25 (2018)
26. Jang, D., Cho, Y., Lee, S., Kim, T., Kim, D.S.: Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. In: ICLR (2025)
27. 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: ECCV. pp. 222–239. Springer (2024)
28. Kang, D., Cho, M.: In defense of lazy visual grounding for open-vocabulary semantic segmentation. In: ECCV. pp. 143–164. Springer (2024)
29. Kim, H., Han, S., Kwon, P., Joo, H.: Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In: ECCV. pp. 400–419. Springer (2024)
30. Kolodiazhnyi, M., Vorontsova, A., Konushin, A., Rukhovich, D.: Oneformer3d: One transformer for unified point cloud segmentation. In: CVPR. pp. 20943–20953 (2024)
31. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
32. Lee, J., Park, C., Choe, J., Wang, Y.C.F., Kautz, J., Cho, M., Choy, C.: Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In: CVPR. pp. 14089–14101 (2025)
33. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: CVPR. pp. 10922–10931 (2023)
34. Li, G., Sun, D., Sevilla-Lara, L., Jampani, V.: One-shot open affordance learning with foundation models. In: CVPR. pp. 3086–3096 (2024)
35. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS* **34**, 19652–19664 (2021)
36. Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al.: What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478* (2024)
37. Li, Y., Zhao, N., Xiao, J., Feng, C., Wang, X., Chua, T.s.: Laso: Language-guided affordance segmentation on 3d object. In: CVPR. pp. 14251–14260 (2024)
38. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)
39. Liu, M., Uy, M.A., Xiang, D., Su, H., Fidler, S., Sharp, N., Gao, J.: Partfield: Learning 3d feature fields for part segmentation and beyond. In: ICCV. pp. 9704–9715 (2025)
40. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: CVPR. pp. 3282–3292 (2022)
41. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning affordance grounding from exocentric images. In: CVPR. pp. 2252–2261 (2022)

42. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Grounded affordance from exocentric view. *IJCV* **132**(6), 1945–1969 (2024)
43. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In: *CVPR*. pp. 16454–16463 (2022)
44. Min, H., Yi, C., Luo, R., Zhu, J., Bi, S.: Affordance research in developmental robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems* **8**(4), 237–255 (2016)
45. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: *ECCV*. pp. 728–755. Springer (2022)
46. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: *CVPR*. pp. 909–918 (2019)
47. Mur-Labadia, L., Guerrero, J.J., Martinez-Cantin, R.: Multi-label affordance mapping from egocentric vision. In: *ICCV*. pp. 5238–5249 (2023)
48. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: *ICRA*. pp. 1374–1381. IEEE (2015)
49. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: *ICCV*. pp. 8688–8697 (2019)
50. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: *IROS*. pp. 5908–5915. IEEE (2017)
51. Nguyen, T., Vu, M.N., Vuong, A., Nguyen, D., Vo, T., Le, N., Nguyen, A.: Open-vocabulary affordance detection in 3d point clouds. In: *IROS*. pp. 5692–5698. IEEE (2023)
52. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *TMLR* (2024)
53. O’Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlikar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: *ICRA*. pp. 6892–6903. IEEE (2024)
54. Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., Li, L.E.: Affordancellm: Grounding affordance from vision language models. In: *CVPRW*. pp. 7587–7597 (2024)
55. Qian, S., Fouhey, D.F.: Understanding 3d object interaction from a single image. In: *ICCV*. pp. 21696–21706 (2023)
56. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *ICML*. pp. 8748–8763. *PmLR* (2021)
57. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: *ICLR*. vol. 2025, pp. 28085–28128 (2025)
58. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: *ICRA*. pp. 8216–8223. IEEE (2023)
59. Tang, J., Wei, Z., Zheng, G., Yang, S.: Closed-loop transfer for weakly-supervised affordance grounding. In: *ICCV*. pp. 9530–9539 (2025)
60. Weder, S., Blum, H., Engelmann, F., Pollefeys, M.: Labelmaker: Automatic semantic label generation from rgb-d trajectories. In: *3DV*. pp. 334–343. IEEE (2024)

61. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS* **35**, 24824–24837 (2022)
62. 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: *ICCV*. pp. 11980–11990 (2025)
63. Xu, C., Chen, Y., Wang, H., Zhu, S.C., Zhu, Y., Huang, S.: Partafford: Part-level affordance discovery from 3d objects (2022)
64. Xu, P., Mu, Y.: Weakly-supervised affordance grounding guided by part-level semantic priors. In: *ICLR* (2025)
65. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. In: *ECCV*. pp. 131–147. Springer (2024)
66. Xue, L., Yu, N., Zhang, S., Panagopoulou, A., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., et al.: Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In: *CVPR*. pp. 27091–27101 (2024)
67. Yang, J., Ding, R., Deng, W., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In: *CVPR*. pp. 19823–19832 (2024)
68. Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., Zha, Z.J.: Grounding 3d object affordance from 2d interactions in images. In: *ICCV*. pp. 10905–10915 (2023)
69. Yang, Y., Huang, Y., Guo, Y.C., Lu, L., Wu, X., Lam, E.Y., Cao, Y.P., Liu, X.: Sampart3d: Segment any part in 3d objects. *arXiv preprint arXiv:2411.07184* (2024)
70. 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: *CVPR*. pp. 1691–1701 (2025)
71. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: *ECCV*. pp. 69–85. Springer (2016)
72. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289* (2023)
73. 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: *CVPR*. pp. 17337–17346 (2025)
74. Zuo, Q., Gu, X., Dong, Y., Zhao, Z., Yuan, W., Qiu, L., Bo, L., Dong, Z.: High-fidelity 3d textured shapes generation by sparse encoding and adversarial decoding. In: *ECCV*. pp. 52–69. Springer (2024)