AFFOGATO: LEARNING OPEN-VOCABULARY AFFORDANCE GROUNDING WITH FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Affordance grounding—localizing object regions based on natural language descriptions of interactions—is a critical challenge for enabling intelligent agents to understand and interact with their environments. However, this task remains challenging due to the need for fine-grained localization, the ambiguity arising from multiple valid interaction regions, and the scarcity of large-scale datasets. We introduce **AffoGato**, a unified framework for open-vocabulary affordance grounding across both 3D and 2D. Our approach leverages supervision from foundation models to automatically generate scalable affordance annotations, enabling training without reliance on exhaustive manual labeling. As part of this pipeline, we construct **Affo-150K**, a large-scale automatically generated dataset of 150K 3D object instances with free-form affordance descriptions and corresponding 3D affordance heatmaps. Within AffoGato, we design simple yet effective models, **Gato-3D** and **Gato-2D**, by combining pre-trained part-aware vision encoders with text-conditional heatmap decoders. Our models achieve state-of-the-art performance across existing 3D and 2D benchmarks, with pretraining on Affo-150K further enhancing their open-vocabulary capabilities.

1 INTRODUCTION

The theory of affordances addresses what an environment offers, provides, or furnishes animals (Gibson, 1979). 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 (Asada et al., 2009; Jamone et al., 2016) and robotics (Chemero & Turvey, 2007; Min et al., 2016; Ardón et al., 2020), here we focus on *visual affordance grounding* from a computer vision perspective (Hassanin et al., 2021); given an object in 3D or 2D vision format, the task is to localize the relevant region for a text description of an interaction with the object. Unlike other types of visual grounding/segmentation (Yu et al., 2016; Kang & Cho, 2024; Lai et al., 2024), the affordance grounding is particularly challenging because an object may have multiple affordances (functionalities) on different regions, the affording regions often have indistinct boundaries, and the affordances are inherently open-vocabulary, i.e., necessitating moving beyond a fixed set of terms for adequate description.

These challenges highlight the need for scalable learning frameworks that can handle a broad spectrum of objects and diverse human-object interactions without relying on exhaustive manual annotation. Existing datasets fall short of this open-ended nature; current 3D datasets (Deng et al., 2021; Nguyen et al., 2023; Yang et al., 2023; Li et al., 2024c; Chu et al., 2025; Yu et al., 2025) are constrained to predefined object and affordance categories (*e.g*. only 23 object classes (Deng et al., 2021)), while 2D datasets cover at most 304 object categories and 36 affordance types (Table 1). Moreover, they contain only predefined affordance labels, limiting generalization to arbitrary text queries or unseen concepts, and their dependence on human annotation restricts scalability.

To overcome these limitations, we propose *AFFOrdance Grounding All aT Once*, dubbed **AffoGato**, a unified learning framework that integrates automatic dataset generation with model training for both 3D and 2D affordance grounding. The key insight is that foundation models (Zhang et al., 2023; Deitke et al., 2025; Gemma Team, 2025) have acquired rich knowledge about object properties and human-object interactions through large-scale pretraining. Our framework leverages this knowledge to automatically synthesize open-vocabulary affordance supervision at scale by enabling fine-grained part-level localization and eliminating the need for manual annotation. Within this framework,

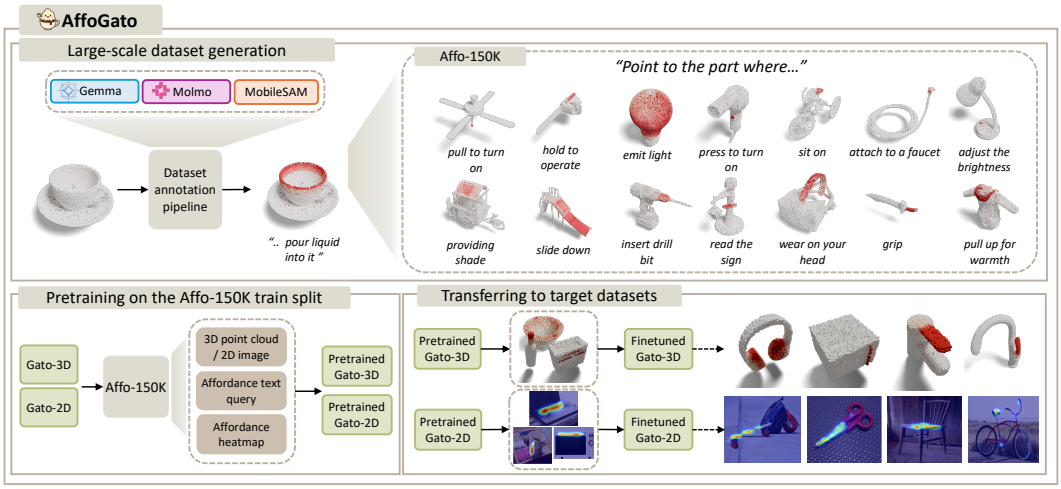

Figure 1: **Overview of AffoGato.** Our framework consists of three stages: large-scale dataset generation, pretraining, and transferring to target datasets. In the dataset generation stage, we leverage foundation models—such as Gemma (Gemma Team, 2025), Molmo (Deitke et al., 2025), and MobileSAM (Zhang et al., 2023)—to construct Affo-150K, a large-scale synthetic dataset comprising 150K 3D objects with open-vocabulary queries and spatially localized heatmap annotations. We pretrain Gato-3D and Gato-2D on the Affo-150K train split, allowing the models to learn generalizable affordance representations. Finally, we finetune our pretrained models on various target datasets, demonstrating strong generalization ability.

we generate **Affo-150K**, the largest and most diverse dataset for affordance grounding to date, comprising 150K 3D object assets from Objaverse (Deitke et al., 2023), each paired with multiple textual affordance descriptions and corresponding 3D heatmaps. Using the obtained 3D heatmaps, we construct a 2D dataset by rendering RGB images of the 3D assets and projecting the aggregated 3D heatmaps into image space, enabling pixel-level supervision aligned with 3D geometry.

We also propose models for both 3D point clouds and 2D images, referred to as **Gato-3D** and **Gato-2D**. Our models share the same architectural concept that differs only in modality-specific vision encoders. When trained on existing benchmarks (Li et al., 2024c; Luo et al., 2022a), they already achieve state-of-the-art performance for both 3D and 2D affordance grounding. When pretrained on Affo-150K, the models demonstrate significantly enhanced performance, particularly improving generalization capabilities to unseen object categories. Despite their simplicity, our models achieve strong performance across both modalities, demonstrating the effectiveness of our framework and the benefits of large-scale pretraining for cross-category generalization. To the best of our knowledge, we are the first to unify the previously separate streams of 3D and 2D affordance grounding within a single scalable framework. To summarize, our contributions are as follows:

- We present AffoGato, a unified framework for 3D and 2D open-vocabulary affordance grounding that automatically generates large-scale training data using foundation models and pretrains affordance grounding models, eliminating manual annotation.
- Our framework generates Affo-150K, the first large-scale synthetic dataset for affordance grounding that leverages foundation models to go beyond predefined affordance and object classes, capturing diverse human-object interactions.
- Extensive experiments show our models significantly outperform existing methods on 3D and 2D benchmarks by large margins, with particularly strong generalization to unseen object categories.

2 Related Work

Visual affordance grounding. Visual affordance grounding aims to localize functional regions in visual data given affordance concepts. Early 2D approaches (Myers et al., 2015; Nguyen et al., 2017; Chuang et al., 2018; Fang et al., 2018; Nagarajan et al., 2019; Luo et al., 2022a; Mur-Labadia et al.,

Table 1: Comparison of 3D (left) and 2D (right) affordance grounding datasets. Aff abbreviates affordances. For Affo-150K, classes and affordances are Gemma3 (Gemma Team, 2025) predictions with frequencies above 10 and 100, respectively.

Dataset	3D Source	# Classes	# Aff	# 3D assets	# Questions
3D AffordanceNet	PartNet	23	18	22,949	0
PartAfford	PartNet	23	24	>25,000	0
OpenAD	3D AffordanceNet	23	37	22,949	0
PIAD	3D AffordanceNet	23	17	7,012	0
LASO	3D AffordanceNet	23	17	8,434	870
SceneFun3D	ARKitScenes	N/A	9	710	17,133
IRAS	OpenAD	23	36	22,949	42,119
SeqAfford	3D AffordanceNet	23	18	18,371	162,386
Affo-150K	Objaverse	**>450**	**>350**	**150,104**	**750,520**

Dataset	# Classes	# Aff	# 2D images
UMD	17	7	10,000
IIT-Aff	10	9	8,835
ADE-Aff	150	3	10,011
ORPA	N/A	7	2,512
Grounded I.H.	31	20	1,871
AGD20K	50	36	23,816
EPIC-Aff	304	20-36	38,876
3DOI	N/A	3	10,000
VRB	N/A	N/A	54,000
Affo-150K	**>450**	**>350**	**150,104**

2023; Qian & Fouhey, 2023; Bahl et al., 2023) focused on predefined affordance categories with limited object diversity. To overcome the scarcity of annotations, recent works explored weakly supervised learning (Luo et al., 2022a), few-shot or zero-shot approaches (Li et al., 2024a; Cuttano et al., 2024), or utilized external videos of human–object interactions (Fang et al., 2018; Nagarajan et al., 2019; Liu et al., 2022; Mur-Labadia et al., 2023; Ju et al., 2024) to produce affordance heatmaps. However, the resultant training data suffer from data noise and biased distribution due to occlusions and limited viewpoints.

In 3D domain, 3D AffordanceNet (Deng et al., 2021) established the first closed-vocabulary 3D affordance grounding benchmark using PartNet (Mo et al., 2019) shapes. PartAfford (Xu et al., 2022) and OpenAD (Nguyen et al., 2023) expanded affordance vocabularies while maintaining categorical constraints. Recent advances toward open-vocabulary 3D grounding include LASO (Li et al., 2024c), which introduced free-form textual queries with 3D heatmap prediction, and 3D AffordanceLLM (Chu et al., 2025), which leveraged language models for query generation. SeqAfford (Yu et al., 2025) focused on textual diversity through GPT-4 generated questions, while SceneFun3D (Delitzas et al., 2024) explored scene-level functionality segmentation.

Despite progress, existing approaches remain constrained by limited training data diversity and scale. Our work addresses these fundamental limitations through automated annotation, enabling robust open-vocabulary affordance grounding across diverse object categories and interaction contexts.

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 (Luo et al., 2022b; Xue et al., 2024; Xu et al., 2024) 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 (Yang et al., 2024b; Liu et al., 2025) distill 2D visual knowledge from foundation models to train 3D geometric encoders, while scene-centric methods (Yang et al., 2024a; Weder et al., 2024; Lee et al., 2025) extend this paradigm to large-scale environments by generating region-level 3D annotations through integration with 2D foundation models. Our work builds upon these ideas to tackle functional understanding via affordance grounding. Instead of focusing on object categorization or part segmentation, we use 2D foundation models on multi-view renderings to produce diverse affordance annotations. This bridges geometric and functional understanding, connecting 2D perception with 3D affordance reasoning.

3 The Affo-150K 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 suffer from significant limitations in scale, diversity, and annotation quality, creating a substantial barrier to progress in this important domain. To address the issue, we propose the *Affo-150K* dataset, a large-scale open-vocabulary affordance grounding dataset. In this section, we elaborate on the pipeline used to generate Affo-150K. Our pipeline leverages large-scale 3D object repositories (Deitke et al., 2023) and state-of-the-art foundation models (Zhang et al., 2023; Gemma Team, 2025; Deitke et al., 2025) to automatically generate high-quality affordance annotations. This addresses the aforementioned limitations by enabling robust and generalizable affordance learning across diverse objects.

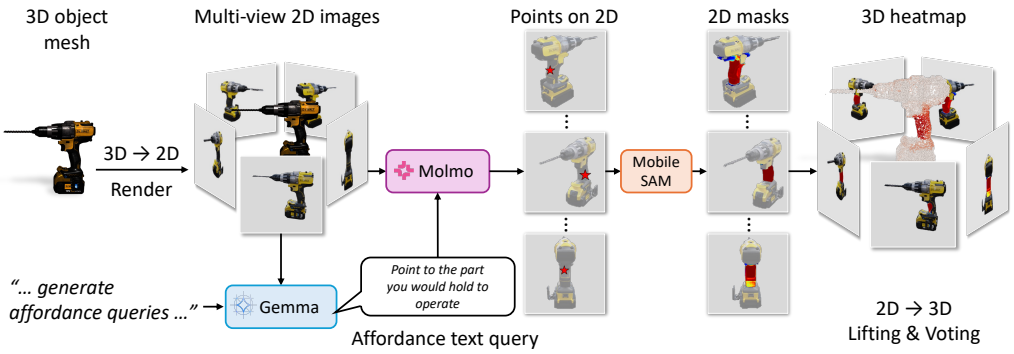

Figure 2: **Overview of our data annotation pipeline.** Given multi-view renderings of an object, Gemma3 (Gemma Team, 2025) generates affordance queries, Molmo (Deitke et al., 2025) points the affordance, and MobileSAM (Zhang et al., 2023) decodes the point to a mask logit. The multi-view mask logits are aggregated on the 3D object surface to obtain an affordance heatmap.

3.1 Source Dataset

We build the Affo-150K dataset upon Objaverse (Deitke et al., 2023), one of the largest public 3D asset repositories. It contains more than 700K web-crawled 3D object meshes spanning diverse functional categories and geometries. Since raw 3D meshes are not directly compatible with recent vision-language models (VLMs) that take 2D images and texts as input, we incorporate G-Objaverse (Zuo et al., 2024), which provides high-resolution, multi-view renderings for over 280K Objaverse objects. These renderings serve as 2D visual input to our annotation pipeline to bridge between 3D objects and language. 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 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. Please refer to Section D for stage-wise qualitative results.

Stage 1. Open-vocabulary affordance query generation. Given multi-view images of an Objaverse object, we employ a multi-modal LLM, Gemma3 (Gemma Team, 2025), to produce natural language queries 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 (Li et al., 2024c; Chu et al., 2025; Yu et al., 2025), 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 capabilities across diverse object geometries and functional categories.

Stage 2. Language-guided interaction point prediction. Once the affordance queries are generated, we utilize Molmo (Deitke et al., 2025), 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

(a) Object classes (b) Affordance classes

(c) Diverse annotations

(d) Heatmap quality

Dataset	Cov. ↑	Div. ↑
LASO (2024c)	0.6384	0.6578
Affo-150K	**0.7532**	**2.6638**

Figure 3: **Characteristics of the Affo-150K dataset.** Please refer to Sec. 3.3 for details.

is used as a prompt for MobileSAM (Zhang et al., 2023) to generate a 2D segmentation mask. We adopt the smallest-mask selection strategy, as it prioritizes precise interactions rather than capturing broader contextual regions. 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 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 we overcome the multi-view inconsistency problem of Molmo and MobileSAM in stage 3, these rendered 2D heatmaps exhibit high consistency across different viewpoints and 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, 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, Affo-150K provides over 150K 3D object instances across four categories: *Daily-Used* includes the largest number of instances (121,799), followed by *Transportations* (11,609), *Furnitures* (8,759), and *Electronics* (7,937). This represents approximately five times the scale of existing datasets (Deng et al., 2021). Each object includes 5 affordance query-heatmap pairs, totaling 750K annotations.

Semantic and spatial diversity. Unlike existing datasets constrained by predefined taxonomies, Affo-150K achieves truly open-vocabulary coverage 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 Affo-150K achieves significantly higher coverage (ratio of points covered by annotation union) and diversity (average pairwise KL divergence between heatmaps) scores than LASO (Li et al., 2024c), indicating richer and more complementary affordance representations.

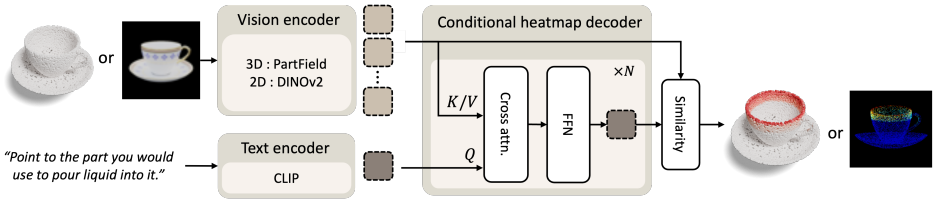

Figure 4: **Gato-3D and Gato-2D architectures.** Each model consists of a 3D or 2D visual encoder, a text encoder, and a text-conditioned heatmap decoder. The affordance heatmap is predicted as the cross-modal similarity of the vision representation and the conditioned text embedding.

Annotation quality evaluation. Beyond diversity and scale, we evaluated the annotation quality to ensure reliability. To verify the efficacy of our automatic annotation pipeline, we conducted a systematic quality assessment where 5K affordance query-heatmap pairs were randomly sampled for human verification. The evaluation demonstrated robust annotation performance, achieving an 84.8% pass rate. Please refer to Section B.1 for detailed statistics on manual evaluation.

Test split construction. We leveraged our quality validation process to construct a rigorous test split. The 5K human-evaluated affordance query-heatmap pairs from the previous section were reserved for testing, ensuring complete train-test separation and evaluation integrity. We engaged human annotators to manually refine the annotations for the subset that fell outside the 84.8% validation threshold. To facilitate this refinement process, we developed a comprehensive manual annotation interface that enables precise editing and validation workflows. The final test split comprises these validated objects with refined annotations, providing a reliable benchmark. Please refer to Section A.1 for details on the manual refinement process.

4 Gato Model Architecture

We present a minimalistic architecture for affordance grounding, dubbed Gato. The architecture is intentionally designed to be simple, yet effective in harnessing the power of the Affo-150K dataset. Building on the shared architectural concept, we create two affordance grounding models: Gato-3D for 3D point clouds and Gato-2D for 2D images. Each model consists of a modality-specific visual encoder, a text encoder, and a text-conditioned heatmap decoder (Fig. 4). The core of our design is a simple text-conditioned heatmap decoder that replaces learnable queries with text embeddings. Although this follows the standard transformer-based mask decoder architecture (Cheng et al., 2021; 2022; Jain et al., 2023; Schult et al., 2023; Kolodiazhnyi et al., 2024), using text embeddings as queries naturally supports open-vocabulary affordance grounding without predefined categories. The following describes the modality-specific designs.

Gato-3D. For the 3D vision encoder, we use the pretrained PartField model (Liu et al., 2025), which captures generic part concepts from large-scale 3D data. For the text encoder, we employ Recap-CLIP (Li et al., 2024b), which provides robust language understanding capabilities. Following PointRefer (Li et al., 2024c), 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. We empirically validate this finetuning strategy in our experiments.

Gato-2D. For the 2D vision encoder, we use the pretrained DINOv2 model (Oquab et al., 2023) as the vision encoder, and CLIP (Radford et al., 2021) as the text encoder. Following OOAL (Li et al., 2024a), we freeze both encoders during training to maintain their pretrained representations. Our architecture adapts the design of OOAL (Li et al., 2024a) 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 evaluate Gato-3D on the LASO dataset (Li et al., 2024c), which supports open-vocabulary 3D affordance grounding with free-form text queries. LASO provides both seen and

Table 2: Open-vocabulary 3D affordance grounding on the LASO (Li et al., 2024c) test split. * denotes models pretrained on the Affo-150K train split and then finetuned on the LASO train split.

Method	Seen				Unseen			
	aIoU↑	AUC↑	SIM↑	MAE↓	aIoU↑	AUC↑	SIM↑	MAE↓
Ref. Trans. (Li & Sigal, 2021)	13.7	79.8	0.497	0.124	10.2	69.1	0.432	0.145
3D-SPS (Luo et al., 2022b)	11.4	76.2	0.433	0.138	7.9	68.8	0.402	0.158
ReLA (Liu et al., 2023)	15.2	78.9	0.532	0.118	10.7	69.7	0.429	0.144
IAGNet (Yang et al., 2023)	17.8	82.3	0.561	0.109	12.9	77.8	0.443	0.129
OpenAD (Nguyen et al., 2023)	14.2	85.1	0.533	0.103	14.6	80.7	0.518	0.109
PointRefer (Li et al., 2024c)	20.8	**87.3**	0.629	**0.093**	14.6	80.2	0.507	0.119
Gato-3D	20.4	86.0	0.633	0.102	18.7	80.0	0.600	**0.101**
Gato-3D*	**21.9**	85.9	**0.637**	0.116	**20.8**	**82.9**	**0.614**	0.122

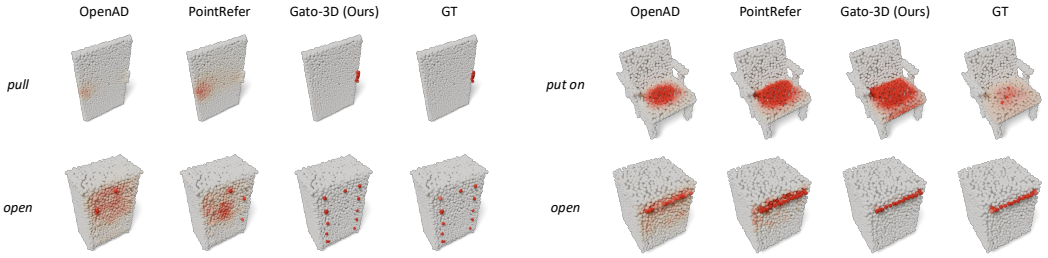

Figure 5: Qualitative comparison between OpenAD (Nguyen et al., 2023), PointRefer (Li et al., 2024c), and Gato-3D on the LASO (Li et al., 2024c) test split.

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.

2D datasets. We evaluate Gato-2D on the AGD20K (Luo et al., 2022a), which consists of two object splits: Seen and Unseen. AGD20K-Weak, 1-shot, and Full represent 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 (Li & Sigal, 2021), 3D-SPS (Luo et al., 2022b), ReLA (Liu et al., 2023), IAGNet (Yang et al., 2023), OpenAD (Nguyen et al., 2023), and PointRefer (Li et al., 2024c). For OpenAD, since official results on LASO (Li et al., 2024c) were not available, we trained the model on the LASO dataset using their official code implementation. All other baseline results are taken directly from Li et al. (2024c). We compare these methods with our proposed Gato-3D on both LASO seen and unseen splits to evaluate generalization performance.

2D baselines. For zero-shot evaluation on AGD20K, we use Molmo+SAM2 (Deitke et al., 2025; Ravi et al., 2024), LISA-7B (Lai et al., 2024) (a reasoning segmentation model), and M^2SA-7B (Jang et al., 2025) (a part-level referring 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 4b, we adopt results from Qian et al. (2024).

Evaluation metrics. For 3D evaluation, following prior work (Li et al., 2024c), 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) (Fiorentino et al., 2023). Higher values are better for aIoU, AUC, SIM, and NSS while lower values are better for MAE and KLD.

5.2 3D AFFORDANCE GROUNDING

LASO (Li et al., 2024c) test split. Table 2 presents the 3D affordance grounding results on LASO. Our proposed Gato-3D achieves competitive performance on the seen setting and demonstrates substantial improvements over existing baselines on the unseen setting. Specifically, Gato-3D

Table 3: Open-vocabulary 3D affordance grounding on the Affo-150K test split. *All → All* shows in-domain results. *Daily-used → Furnitures* and *Furnitures → Daily-used* show cross-domain results where models are trained on one category and evaluated on the other.

Method	*All → All*				*Daily-used → Furnitures*				*Furnitures → Daily-used*			
	aIoU↑	AUC↑	SIM↑	MAE↓	aIoU↑	AUC↑	SIM↑	MAE↓	aIoU↑	AUC↑	SIM↑	MAE↓
OpenAD	3.1	64.8	0.329	0.150	7.6	67.8	0.368	0.177	1.6	54.3	**0.308**	0.201
PointRefer	10.5	76.1	0.405	0.120	13.5	78.9	0.443	0.157	2.8	60.6	0.265	**0.122**
Gato-3D	**13.6**	**79.0**	**0.429**	**0.111**	**18.2**	**80.4**	**0.475**	**0.125**	**4.6**	**62.4**	0.304	0.130

Table 4: Open-vocabulary 2D affordance grounding on the AGD20K (Luo et al., 2022a) test split. For zero-shot evaluation, Gato-2D is pretrained on the Affo-150K train split. * denotes models pretrained on the Affo-150K train split and then finetuned on the AGD20K train split with full supervision.

(a) Zero-shot results

Method	KLD↓	SIM↑	NSS↑
Seen split			
Molmo+SAM2 (2025; 2024)	1.804	0.261	0.729
LISA-7B (2024)	1.627	0.296	0.819
M^{2}SA-7B (2025)	1.772	0.258	0.620
Gato-2D	**1.426**	**0.402**	**0.985**
Unseen split			
Molmo+SAM2 (2025; 2024)	1.953	0.226	0.718
LISA-7B (2024)	1.830	0.256	0.765
M^{2}SA-7B (2025)	1.925	0.227	0.657
Gato-2D	**1.571**	**0.376**	**1.016**

(b) Supervised learning results

Method	Sup.	KLD↓	SIM↑	NSS↑
Cross-view-AG (2022a)		1.787	0.285	0.829
Cross-view-AG+ (2024)		1.765	0.279	0.882
AffCorrs (2023)	Weak	1.618	0.348	1.021
LOCATE (2023)		1.405	0.372	1.157
WSAG-PLSP (2025)		1.153	0.437	1.418
OOAL (2024a)	1-shot	1.070	0.461	1.503
LOCATE-Sup (2023)		1.907	0.236	0.641
LOCATE-Sup-OWL (2023; 2022)		1.927	0.234	0.624
AffordanceLLM (2024)	Full	1.463	0.377	1.070
Gato-2D		1.034	0.503	1.550
Gato-2D*		**0.974**	**0.519**	**1.645**

achieves competitive results compared to the previous SOTA method PointRefer (Li et al., 2024c) on the seen setting (20.4 vs. 20.8 aIoU) and shows significant gains by a large margin of 4.1%p aIoU (18.7 vs. 14.6) on the unseen setting, outperforming all existing methods. The effectiveness of Affo-150K pretraining is particularly evident when finetuned on LASO, where Gato-3D* achieves the best performance on both seen and unseen settings. Compared to the non-pretrained version, the pretrained model shows improvements of 1.5%p aIoU (21.9 vs. 20.4) on the seen setting and more substantial gains of 2.1%p aIoU (20.8 vs. 18.7) on the unseen setting. The pretraining effect is especially pronounced on the unseen setting, demonstrating the value of large-scale pretraining for generalization to novel object-affordance combinations. Figure 5 provides qualitative comparison showing Gato-3D's superior affordance localization on the LASO test split.

Affo-150K test split. We evaluate the open-vocabulary generalization ability of baseline models on Affo-150K using both in-domain and cross-domain settings (Tab. 3). We train and evaluate OpenAD (Nguyen et al., 2023) and PointRefer (Li et al., 2024c) using their official code implementations on our dataset. For in-domain evaluation (*All → All*), Gato-3D achieves the best performance across all metrics, with an aIoU of 13.6, AUC of 79.0, and SIM of 0.429, demonstrating superior affordance grounding capabilities. For cross-domain evaluation, we train models on either the *Daily-Used* or *Furnitures* subset and evaluate on the other, requiring both domain and category generalization. Gato-3D consistently outperforms prior methods in both directions: when trained on *Daily-Used* and tested on *Furnitures*, it achieves an aIoU of 18.2 and AUC of 80.4; when trained on *Furnitures* and tested on *Daily-Used*, it achieves an aIoU of 4.6 and AUC of 62.4. This demonstrates Gato-3D's strong ability to generalize affordance concepts to unseen object categories, highlighting its effective cross-domain generalization in open-vocabulary settings.

5.3 2D AFFORDANCE GROUNDING

Table 4a demonstrates the exceptional zero-shot generalization capability of Gato-2D on AGD20K. Despite being a lightweight combination of DINOv2-ViT/B image encoder and CLIP text encoder, our model significantly outperforms heavily parameterized LLM-based approaches across both seen and unseen object splits. This superior generalization performance highlights the effectiveness of large-scale Affo-150K pretraining for learning robust affordance representations that transfer well to new domains and object categories. Please refer to section Section D for visualizations.

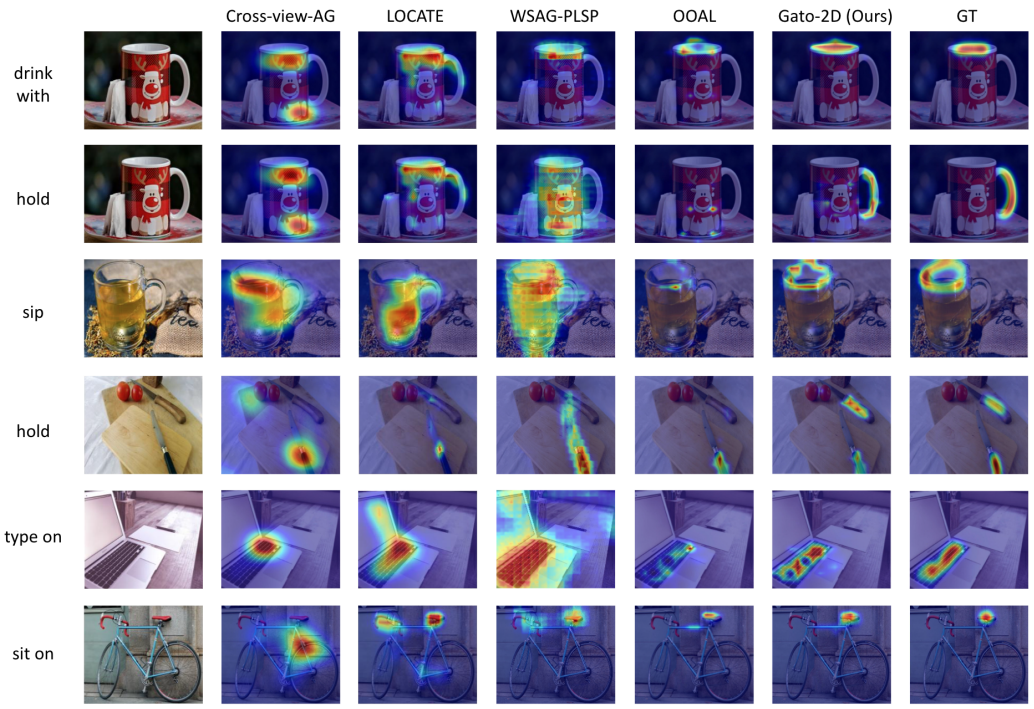

Figure 6: Qualitative comparison between Cross-view-AG (Luo et al., 2022a), LOCATE (Li et al., 2023), WSAG-PLSP (Xu & Yadong, 2025), OOAL (Li et al., 2024a), and Gato-2D, which is pretrained on the Affo-150K train split and then finetuned on the AGD20K train split with full supervision.

Table 4b presents supervised learning results across different supervision levels. Gato-2D pretrained on Affo-150K and fine-tuned with full AGD20K supervision achieves state-of-the-art performance. Notably, Affo-150K pretraining provides consistent improvements over training solely on AGD20K, demonstrating the value of large-scale pretraining even when fine-tuning data is available. This validates Affo-150K's effectiveness as a pretraining dataset that enhances performance despite domain differences, thanks to its diversity and scale. Figure 6 shows qualitative comparison showing Gato-2D's superior performance.

6 Discussion and Conclusion

We have presented AffoGato, a unified framework for open-vocabulary affordance grounding across both 3D and 2D domains. By leveraging foundation models, our approach enables automatic generation of large-scale supervision signals, eliminating the need for manual annotation and capturing diverse human-object interactions beyond predefined categories. Our approach produces Affo-150K, the largest dataset for affordance grounding with 150K diverse 3D assets and free-form affordance descriptions, moving beyond the predefined categories that have limited existing work. We also propose Gato-3D and Gato-2D models that share a unified architectural concept with modality-specific vision encoders. Despite their simplicity, these models achieve state-of-the-art performance on existing benchmarks with particularly strong generalization to unseen object categories, demonstrating the effectiveness of large-scale pretraining on Affo-150K. While our approach facilitates scalable learning of contact points related to an action, an embodied agent in the real world requires predicting more detailed and fine-grained information beyond spatial localization, *e.g.*, temporal dynamics and force requirements (Kim et al., 2024), which is limited in our current work. We believe that AffoGato framework and the Affo-150K dataset will provide a solid foundation for future research toward more comprehensive embodied AI systems.

7 Reproducibility Statement

To ensure the reproducibility of our work, we provide comprehensive implementation details and resources throughout the paper and supplementary materials. In Sec. A, we detail our data annotation pipeline, specifying the exact foundation models used, computational requirements including GPU usage and processing time, and provide complete architectural details for our Gato-3D and Gato-2D models including loss functions, batch sizes, and training configurations. We include an interactive annotation interface screenshot in Fig. 7 to illustrate our human evaluation methodology. Additionally, we visualize intermediate results of our annotation pipeline in Fig. 12 to help others reproduce our data generation process. Finally, we have officially released our Affo-150K dataset through Hugging Face at https://huggingface.co/datasets/project-affogato/affogato, making our 150K annotated instances publicly available for future research and comparison.

References

Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020.

Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. Cognitive developmental robotics: A survey. *IEEE transactions on autonomous mental development*, 1(1):12–34, 2009.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13778–13790, 2023.

Anthony Chemero and Michael T Turvey. Gibsonian affordances for roboticists. *Adaptive Behavior*, 15(4):473–480, 2007.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Hengshuo Chu, Xiang Deng, Qi Lv, Xiaoyang Chen, Yinchuan Li, Jianye HAO, and Liqiang Nie. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 975–983, 2018.

Claudia Cuttano, Gabriele Rosi, Gabriele Trivigno, and Giuseppe Averta. What does clip know about peeling a banana? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2238–2247, 2024.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. In *CVPR*, 2025.

Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14531–14542, 2024.

Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1778–1787, 2021.

Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2139–2147, 2018.

Maria Chiara Fiorentino, Francesca Pia Villani, Mariachiara Di Cosmo, Emanuele Frontoni, and Sara Moccia. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Medical image analysis*, 83:102629, 2023.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 1979.

Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th international conference on computer vision*, pp. 1–8. IEEE, 2009.

Denis Hadjivelichkov, Sicelukwanda Zwane, Lourdes Agapito, Marc Peter Deisenroth, and Dimitrios Kanoulas. One-shot transfer of affordance regions? affcorrs! In *Conference on Robot Learning*, pp. 550–560. PMLR, 2023.

Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2989–2998, 2023.

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25, 2016.

Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Dae-Shik Kim. Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. *ICLR*, 2025.

Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, pp. 222–239. Springer, 2024.

Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 143–164. Springer, 2024.

Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *European Conference on Computer Vision*, pp. 400–419. Springer, 2024.

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20943–20953, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Junha Lee, Chunghyun Park, Jaesung Choe, Yu-Chiang Frank Wang, Jan Kautz, Minsu Cho, and Chris Choy. Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10922–10931, 2023.

Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3086–3096, 2024a.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024b.

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

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23592–23601, 2023.

Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. In *ICCV*, 2025.

Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3282–3292, 2022.

Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in neural information processing systems*, 32, 2019.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2252–2261, 2022a.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *International Journal of Computer Vision*, 132(6):1945–1969, 2024.

Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16454–16463, 2022b.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Huaqing Min, Chang'an Yi, Ronghua Luo, Jinhui Zhu, and Sheng Bi. Affordance research in developmental robotics: A survey. *IEEE Transactions on Cognitive and developmental systems*, 8 (4):237–255, 2016.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pp. 728–755. Springer, 2022.

Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 909–918, 2019.

Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5238–5249, 2023.

Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pp. 1374–1381. IEEE, 2015.

Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8688–8697, 2019.

Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5908–5915. IEEE, 2017.

Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5692–5698. IEEE, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21753–21763, 2023.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7587–7597, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8216–8223. IEEE, 2023.

Silvan Weder, Hermann Blum, Francis Engelmann, and Marc Pollefeys. Labelmaker: automatic semantic label generation from rgb-d trajectories. In *2024 International Conference on 3D Vision (3DV)*, pp. 334–343. IEEE, 2024.

Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. In *ECCV VOLI Workshop*, 2022.

Peiran Xu and MU Yadong. Weakly-supervised affordance grounding guided by part-level semantic priors. In *The Thirteenth International Conference on Learning Representations*, 2025.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2024.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19823–19832, 2024a.

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

Yunhan Yang, Yukun Huang, Yuan-Chen Guo, Liangjun Lu, Xiaoyang Wu, Edmund Y Lam, Yan-Pei Cao, and Xihui Liu. Sampart3d: Segment any part in 3d objects. *CoRR*, 2024b.

Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. Seqaffordance: Sequential 3d affordance reasoning via multimodal large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

Qi Zuo, Xiaodong Gu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Lingteng Qiu, Liefeng Bo, and Zilong Dong. High-fidelity 3d textured shapes generation by sparse encoding and adversarial decoding. In *European Conference on Computer Vision*, 2024.

A IMPLEMENTATION DETAILS

A.1 DATA ANNOTATION PIPELINE

Use of pretrained models. Our data annotation pipeline uses the following pretrained models: (1) Gemma3 (Gemma Team, 2025) (`google/gemma-3-4b-it`) for generating natural language affordance queries (2) Molmo (Deitke et al., 2025) (`allenai/Molmo-7B-D-0924`) for predicting interaction points in 2D images, and (3) MobileSAM (Zhang et al., 2023) for predicting 2D heatmap given interaction point prompts.

Computational resources. For Affo-150K dataset generation, we use 8 NVIDIA H100 GPUs with 80GB of memory each. The data generation pipeline takes approximately 24 hours to process 150K Objaverse instances.

Image sampling. G-Objaverse (Zuo et al., 2024) provides 38 views per object. For computational efficiency, we used the first 25 views that are captured with the same elevation but uniformly distributed azimuths around the object. For stage 1 of our pipeline, we sample 5 images at equal intervals from these 25 views to generate affordance queries. This sampling strategy ensures comprehensive coverage of the object from multiple perspectives while optimizing computational resources. For the remaining stages, we utilized all 25 views.

Human evaluation. As discussed in Section 3.3, we instruct human annotators to evaluate the quality of our automatically generated annotation and to refine annotations for those didn't pass the quality check. We guide the annotators to rate the affordance query-heatmap pairs from three criteria: (1) semantic relevance between the query and object, (2) spatial accuracy of the predicted interaction points, and (3) coverage of the heatmap for the intended affordance. We provide the annotators with a web-based interactive viewer for screening the affordance query-heatmap pairs and assigning ratings based on the three criteria. The example of the web-based interactive UI is shown in Figure 7.

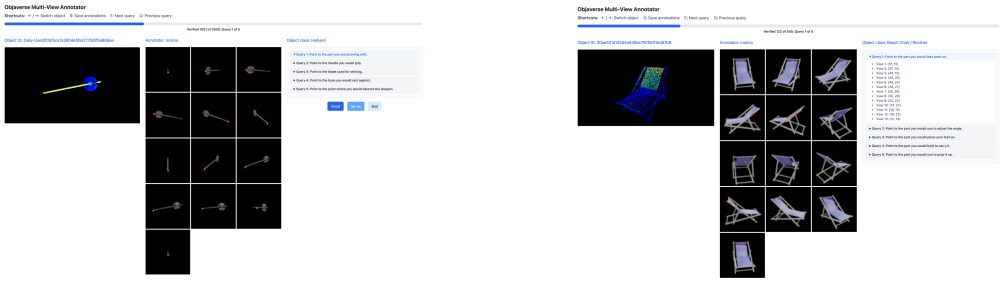

Figure 7: Web-based interactive viewer for (Left) quality evaluation and (Right) human refinement

A.2 3D AFFORDANCE GROUNDING

Gato-3D architecture. Our Gato-3D model leverages the PartField (Liu et al., 2025) architecture for processing 3D visual inputs and incorporates Recap-CLIP (Li et al., 2024b) for encoding language queries. Specifically, we adopt a hybrid 3D encoder composed of PVCNN (Liu et al., 2019) and a triplane transformer, as introduced in the official PartField repository. We used the official pretrained checkpoint and froze the entire vision backbone during training to ensure consistent feature extraction.

For the text branch, we utilize the Recap-CLIP text encoder, which provides enhanced language grounding compared to standard CLIP variants. The resulting query embeddings are fed into a conditional heatmap decoder that predicts spatial affordance distributions over 3D points. The decoder augments the 3D point features from the vision encoder with Fourier-based positional encodings and uses them as keys and values in a cross-attention mechanism, where the language embeddings act as queries. The attended features are then refined through a residual feedforward network (an MLP with skip connections), which outputs the final heatmap over the point cloud.

Training detail. To ensure a fair comparison with prior methods, we adopt the same loss formulation as LASO (Li et al., 2024c). The model is optimized using a combination of Binary Cross-Entropy (BCE) Loss to handle classification and Dice Loss (Milletari et al., 2016) to improve region-level alignment. The two losses are summed with equal weights to form the final training objective. We

Table 5: Effect of random background augmentation

(a) Zero-shot evaluation on AGD20K (Luo et al., 2022a)

Method	Affo-150K pretrain	Data augmentation	KLD ↓	SIM ↑	NSS ↑
Seen split					
Gato-2D	✓		1.493	0.355	0.920
Gato-2D	✓	✓	**1.426**	**0.402**	**0.985**
Unseen split					
Gato-2D	✓		1.688	0.313	0.876
Gato-2D	✓	✓	**1.571**	**0.376**	**1.016**

(b) Fine-tuning on the AGD20K-Full

Method	Affo-150K pretrain	Data Augmentation	KLD↓	SIM↑	NSS↑
Gato-2D	✓		1.100	0.470	1.497
Gato-2D	✓	✓	**0.974**	**0.519**	**1.645**

use the same training setup for both the LASO and Affo-150K datasets: 50 epochs, batch size of 64, and training on 8 NVIDIA RTX A6000 GPUs.

A.3 2D AFFORDANCE GROUNDING

AGD20K dataset. AGD20K-Weak refers to the original AGD20K dataset. The training set consists of 23,083 / 13,323 image-level labels for the Seen / Unseen splits, respectively, while the corresponding test sets contain 1,675 / 540 images. AGD20K-Oneshot refers to the AGD20K dataset for one-shot affordance learning. The training set consists of 50 / 33 images—one per object class—for the Seen / Unseen splits, respectively. The test set is identical to that of AGD20K-Weak. AGD20K-Full is constructed for fully supervised training, following the setup of Qian et al. (2024). The training set consists of 999 images including object classes from the training set of AGD20K's unseen split, each annotated with dense pixel-level affordance masks. The test set contains 540 images from object classes in the test set of the unseen split.

Background augmentation in Gato-2D pretraining stage. For the pretraining stage of Gato-2D, we replace the null background in each rendered image with a randomly selected background from the Background dataset (Gould et al., 2009). After background replacement, the image is resized to 256×256, randomly cropped to 224×224, and horizontally flipped with a random probability. Table 5 summarizes zero-shot performance on AGD20K with and without background augmentation during pre-training, as well as the fine-tuning results on AGD20K-Full. Empirically, we observe that background augmentation leads to better generalization compared to pretraining without it.

Gato-2D architecture. Our Gato-2D architecture is adapted from OOAL (Li et al., 2024a). Multi-level features from different layers of DINOv2 are aggregated. To focus attention on foreground regions, cross-attention is restricted to the regions indicated by the mask derived from the CLS token. Unlike OOAL, which employs text prompt learning with fixed affordance labels as input, our model takes a single natural language query as the text input without using any text prompt learning.

Training detail. We used the CLIP ViT-B/16 as the text encoder and DINOv2 ViT-B/14 as the vision backbone. During pre-training on Affo-150K, the model is optimized using Adam with a learning rate of 0.001. The training is conducted for 52,000 iterations with a per-GPU batch size of 512 on 7 NVIDIA RTX 3090 GPUs. For fine-tuning on AGD20K-Full, we use the Adam optimizer with a learning rate of 0.0001. Training is performed for 400 iterations with a batch size of 512 on a single NVIDIA RTX 3090 GPU. Binary cross-entropy loss is employed consistently in both the pretraining and fine-tuning stages.

Figure 8: Effect of CoT Prompting on Affordance Query Generation.

Method	Training Res.	Test Resolution (# Points)			
	(# Points)	**2048**	**4096**	**8192**	**16384**
Gato-3D	2048	0.237	0.245	0.249	0.251
	16384	0.265	0.278	0.283	0.287

Table 6: **Performance of Gato-3D under varying training and test resolutions.** When the training and test resolutions differ, we first perform inference at the training resolution and then interpolate the predicted heatmaps to match the target test resolution.

B Analyses

B.1 Analyses on Data Annotation Pipeline

CoT alleviates affordance misprediction. We observe that CoT prompting helps reduce affordance prediction errors. As shown in Figure 8, without CoT prompting, the model fails to capture the object's functional properties and instead relies primarily on its shape when generating affordance queries. 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, with CoT prompting, the model infers functionally meaningful outputs such as "Point to the handle to open the door," or "Point to the part you would place the hat on."

Effect of data resolution. Affo-150K provides point clouds with a high resolution of 16,384 points, enabling the capture of fine-grained geometric details critical for affordance understanding. This represents a significant improvement over LASO (Li et al., 2024c), which provides only 2,048 points resolution. To assess the effect of data resolution, we trained Gato-3D on Affo-150K at two resolutions (2,048 vs. 16,384 points) and report the results in Table 6. The model trained with high-resolution data consistently outperforms its low-resolution counterpart across all test resolutions, indicating that learning fine-grained geometric details during training is crucial for accurate 3D affordance grounding at any scale.

Human vs. our annotation pipeline. We compare the affordance predictions from our annotation pipeline with human annotations on 3D-AffordanceNet (Bahl et al., 2023) meshes. Figure 9 presents a qualitative comparison, where the first and third columns display affordance heatmaps generated by our annotation pipeline, while the second and fourth columns show human-annotated ground truth from 3D-AffordanceNet. The visual comparison demonstrates that our automated pipeline produces affordance heatmap predictions that closely align with human intuition about object affordances. This suggests that our annotation pipeline can serve as a reliable substitute for manual annotation, significantly reducing the time and effort required to create large-scale datasets.

Failure modes of annotation pipeline. Despite enabling automatic affordance annotation at scale, our annotation pipeline exhibits several failure modes in its pipeline. First, despite our use of Chain-of-Thought prompting to mitigate errors, Gemma3 (Gemma Team, 2025) occasionally produces affordance queries that are semantically misaligned with the object's functionality. This occurs when the LLM misidentifies the object category, which then propagates through the chain-of-thought prompting to generate irrelevant affordance queries. Second, SAM (Zhang et al., 2023; Ravi et al., 2024) tends to be biased towards object edges, leading to heatmap predictions that over-emphasize boundaries rather than functionally relevant regions. This edge bias can result in incomplete or imprecise affordance annotations, particularly for affordances that involve interacting with the interior regions of objects. Third, there are cases where the target object part described in the affordance query

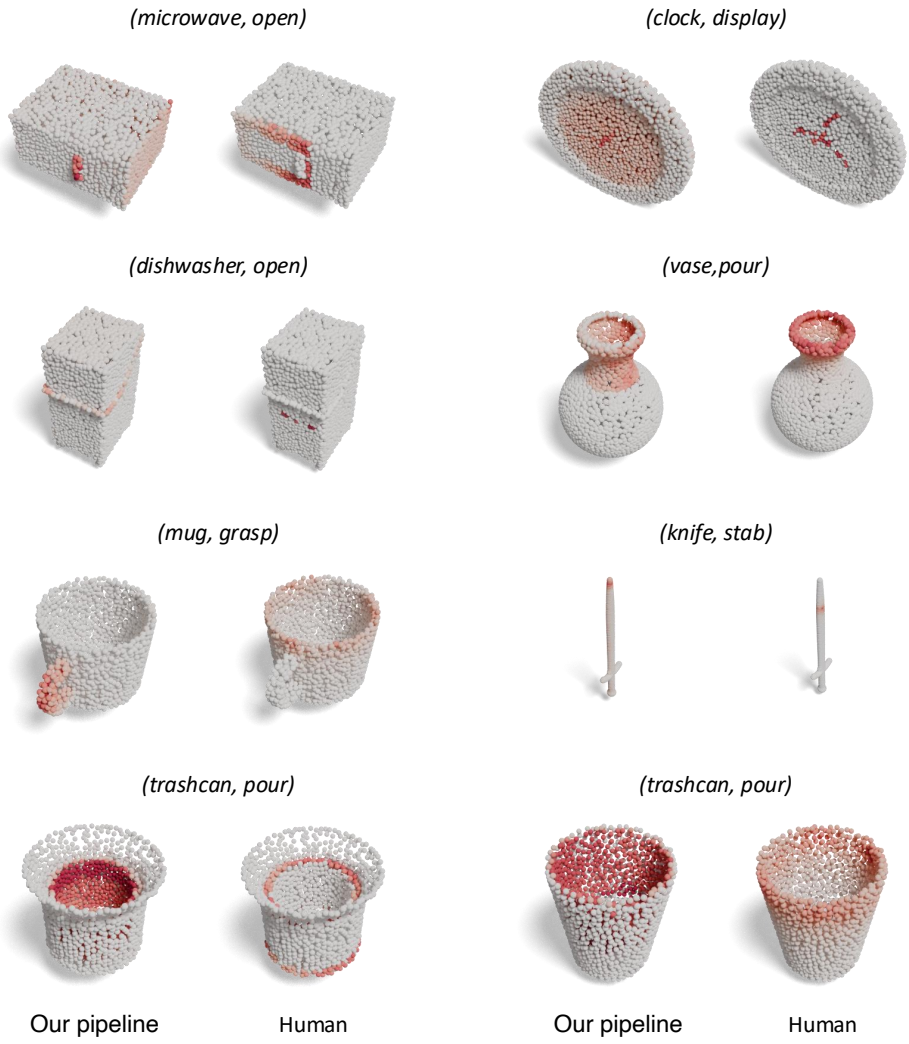

Figure 9: **Our annotation pipeline vs. Human**. Comparison is conducted on 3D-AffordanceNet (Bahl et al., 2023) meshes. First and third columns show affordance heatmaps predicted by Affo-150K annotation pipeline, while second and fourth columns are human-annotated.

is not visible in the multi-view images due to occlusion or camera angle limitations. For example, given the affordance query *"Point to the part where you would use to brake the car"*, the brake pedal is often occluded since our multi-view images are captured using a circular camera trajectory around the object's exterior. To ensure dataset quality, we address these visibility issues through our human evaluation process, which filters out such problematic cases from the test split. The distribution of failure modes across categories is shown in Figure 10.

B.2 Failure cases of Gato

Figure 11 illustrates failure cases from the zero-shot evaluation of the Gato-2D model, pretrained on Affo-150K, on the AGD20K dataset. Images in AGD20K sometimes contain multiple objects, which can result in several plausible regions corresponding to a single affordance. For example, in the left image, the "hold" affordance could refer either to the handle of the coffee machine or to the handle of the cup. Similarly, in the right image, "hold" could apply to either the chopsticks or the outer surface of the bowl. These cases highlight limitations in the ground truth annotations. To make more precise predictions, it may be necessary to include explicit object information in the prompt, such as "hold

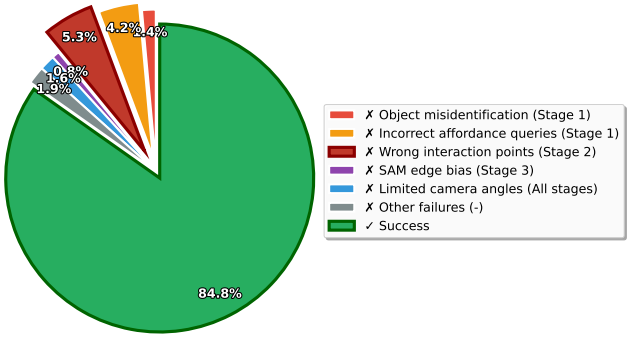

Figure 10: Annotation quality evaluation results from Section 3.3.

Figure 11: Failure cases of Gato-2D in zero-shot evaluation on AGD20K.

the cup" or "hold the bowl." Our model is capable of accepting arbitrary natural language queries, making it well-suited for resolving such disambiguities effectively.

C LIMITATIONS

As the Affo-150K dataset is derived from the Objaverse 3D assets, our data do not contain the background information. Due to this limitation, we randomly synthesize background on the 2D images as shown in Table 5, which is shown to be helpful when transferred to the real-world images. Note that our data engine can be extended to the indoor or outdoor scene data to tackle navigation environments, leaving them for future work.

D ADDITIONAL QUALITATIVE RESULTS

Stage-wise qualitative results. Figure 12 presents the intermediate qualitative results at each stage of our pipeline. Stage 1 takes multi-view images as input and predicts the object's class name together with five candidate queries. Stage 2 leverages Molmo to point to the affordance region corresponding to each query. Stage 3 converts Molmo's pointing locations into pixel-wise masks. These results highlight the robustness of our pipeline: even when Molmo produces incorrect pointings for some views or MobileSAM generates imprecise masks, the multi-view consensus voting effectively suppresses such errors, yielding an accurate final output.

Qualitative comparison of zero-shot evaluation on AGD20K. In Figure 13, we visualize qualitative results for 2D zero-shot affordance grounding. It illustrates common failure modes of existing methods (Lai et al., 2024; Jang et al., 2025) that capture whole objects rather than precise part-level affordances. While Molmo+SAM provides rough part-level localization, our model trained on the multi-view aggregated Affo-150K dataset achieves refined grounding capabilities.

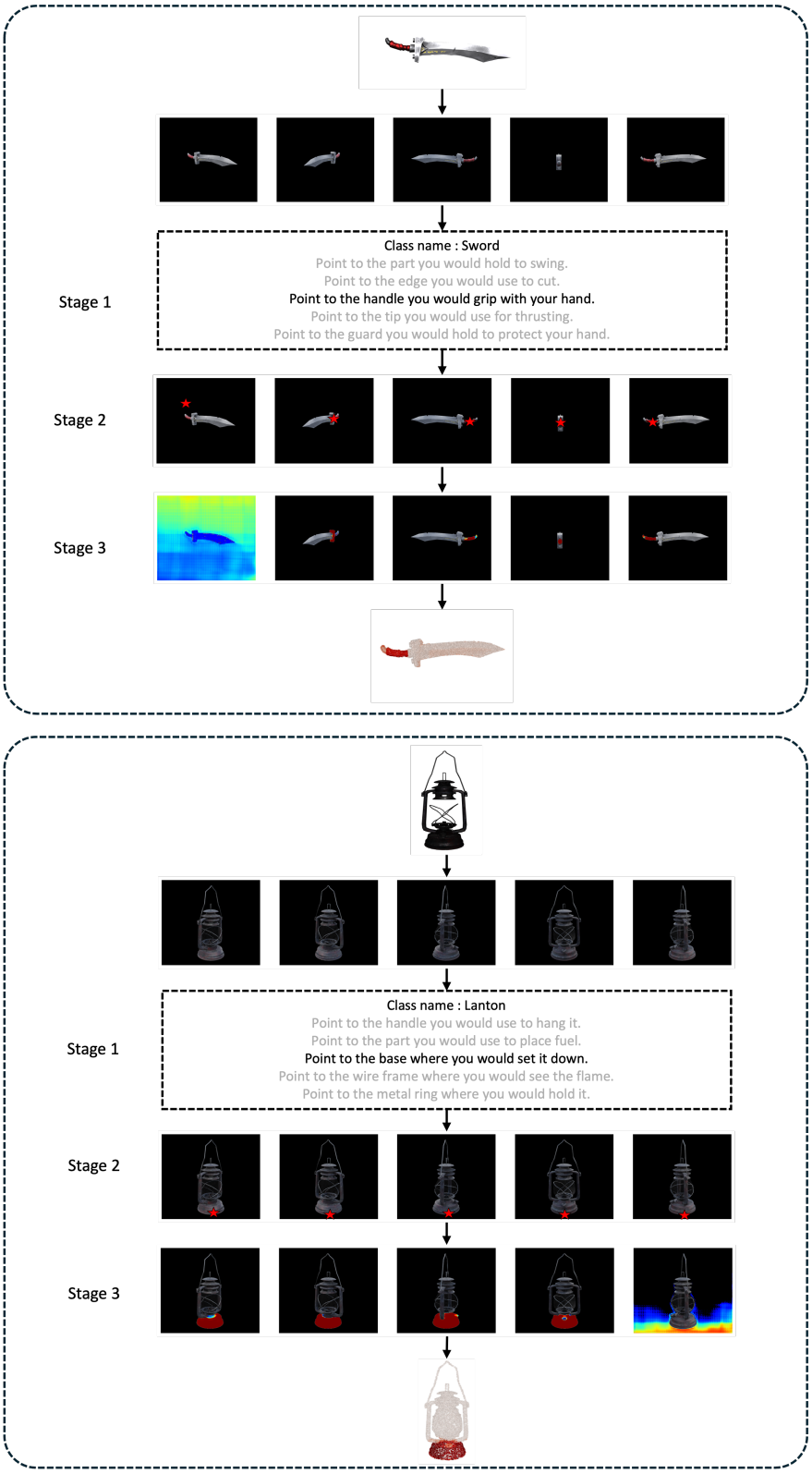

Figure 12: **Intermediate qualitative results at each stage of annotation pipeline.** For Stage 2 and Stage 3, we present the results corresponding to the affordance query highlighted in black from Stage 1. In Stages 2 and 3, the visualizations are shown for five representative views, uniformly sampled from the all 25 views.

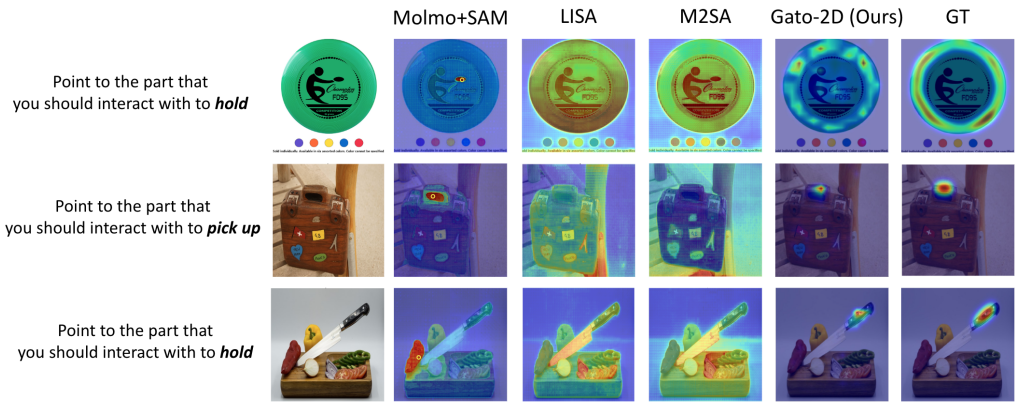

Figure 13: Zero-shot evaluation comparison of Gato-2D and others on AGD20K (Luo et al., 2022a)