AffordMatcher: Affordance Learning in 3D Scenes from Visual Signifiers

Nghia Vu[†,2], Tuong Do[†,1,2,3], Khang Nguyen[4], Baoru Huang[1,7,*], Nhat Le[5], Binh Xuan Nguyen[2], Erman Tjiputra[2], Quang D. Tran[1,2], Ravi Prakash[6], Te-Chuan Chiu[3], Anh Nguyen[1]

[1]University of Liverpool, UK [2]AIOZ Ltd., Singapore [3]National Tsing Hua University, Taiwan

[4]MBZUAI [5]University of Western Australia [6]Indian Institute of Science [7]NVIDIA

https://aioz-ai.github.io/AffordMatcher/

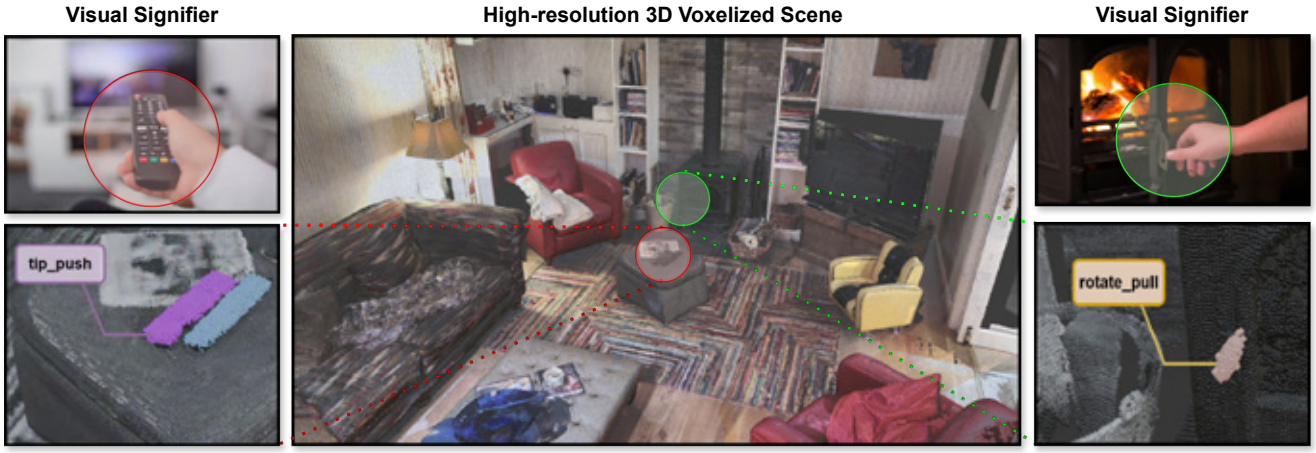


Figure 1. Overview of *AffordMatcher*: Detecting and localizing affordances in 3D voxelized scenes through visual signifiers entails semantic context drawn from RGB images. Given a scene representation and visual signifiers, *AffordMatcher* can understand actionable commands, such as "*watch the television*", "*push the tip*", "*rotate pull*", or "*open the chimney*", and identify spatial affordances.

Abstract

Affordance learning is a complex challenge in many applications, where existing approaches primarily focus on the geometric structures, visual knowledge, and affordance labels of objects to determine interactable regions. However, extending this learning capability to a scene is significantly more complicated, as incorporating object- and scene-level semantics is not straightforward. In this work, we introduce AffordBridge, a large-scale dataset with $291,637$ *functional interaction annotations across* 685 *high-resolution indoor scenes in the form of point clouds. Our affordance annotations are complemented by RGB images that are linked to the same instances within the scenes. Building upon our dataset, we propose AffordMatcher, an affordance learning method that establishes coherent semantic correspondences between image-based and point cloud-based instances for keypoint matching, enabling a more precise identification of affordance regions based on cues, so-called visual signifiers. Experimental results on our dataset demonstrate the effectiveness of our approach compared to other methods.*

[†] equal contribution; * corresponding author

1. Introduction

Humans interact with the environment as part of their daily routines. Analyzing these interactions can provide valuable insights and is highly beneficial for both humans and robots to perform meaningful actions in the given environment. To better understand the interaction between humans and the environment, Gibson introduced the concept of "*affordance*", referring to "*opportunities for interaction*" [18]. Yet, these opportunities need to be physically verifiable through successful actions to determine whether they are true *signifiers*; one point was later discussed by Norman [49]. Therefore, learning about affordances requires not only predicting types of interactions but also correctly identifying specific points on objects that facilitate these human-object interactions. The concept of affordance from signifiers brings together perception and action, thus opening up applications in robotic manipulation [1, 6, 71, 72], human-robot interaction [2, 77], visual navigation [25, 41], and augmented reality [53, 75].

Still, the concept of "*affordance*" standalone is broad. Depending on the context, affordance learning is typically treated as a single-modality learner. Image-based affordance learning predicts the corresponding segmentation maps at the

Dataset / Attribute	Total Samples	Environment	Form of Interactions		Affordance Annotation	No. Affordances	No. Categories	No. Aff. Actions
			Implicit	*Explicit*				
EPIC-Aff [42]	38,876	2D Images	–	–	2D masks	–	304	43
AGD20k [35]	23,816	2D Images	–	–	2D boxes	–	50	36
PartNet [39]	26,671	3D Objects	–	–	3D masks	573,585	–	24
AffordPose [21]	641	3D Objects	–	Grasp hand poses	3D masks	26,712	13	8
3DAffordanceNet [14]	22,949	3D Objects	–	–	3D masks	56,307	23	18
PIAD [74]	7,012	3D Objects	Single interactions	–	3D masks	7,012	23	17
LASO [30]	8,434	3D Objects	Commands on objects	–	3D masks	19,751	23	17
Scenefun3D [9]	–	3D Scenes	–	Commands on scenes	3D masks	14,279	–	9
PIADv2 [56]	38,889	3D Objects	Single interaction	–	3D masks	38,889	43	24
AED [26]	–	2D Images	–	–	2D masks	–	13	8
SeqAfford [76]	18,371	3D Objects	Commands on objects	–	3D masks	183,233	23	18
MIPA [17]	7,012	3D Scenes	Multiple interactions	–	3D masks	7,012	23	17
***AffordBridge* (Ours)**	317,844	3D Scenes	Visual signifiers	Descriptions	3D masks	291,637	157	61

Table 1. Comparisons between *AffordBridge* and other datasets: Our dataset introduces a large-scale benchmark for spatial affordance identification from visual signifiers through both implicit and explicit human-object interactions. *AffordBridge* contains $317,844$ high-resolution paired samples of 2D-3D representations across 685 scenes. The interacted objects are annotated by 3D masks, yielding $291,637$ volumetric masks of interactable regions among 157 object categories through 61 actionable affordances.

pixel level for intended actions [14, 34, 43, 54]. Meanwhile, spatial affordance learning segments desired affordance masks on object point clouds at the voxel level [10, 47, 67]. The share of the two mentioned approaches leverages text prompts to direct affordance learning. However, the fusion of these two modalities remains unclear [36, 45]. Through this, the lack of multimodal representation learning for affordance conveys the idea of imposing cross-modal learning.

In practice, many challenges coexist in localizing spatial affordance from visual signifiers. First, cross-modal representations require overcoming significant discrepancies in feature distributions between images and point clouds [11]. Second, matching affordance localization in the 3D domain and affordance detection in image space under diverse actions across different scenes entails the dexterity of designing such learning models [11]. Additionally, language instructions, such as "*press here*" or "*rotate knob*", are indeed semantically ambiguous without explicit geometric context [74], not to mention that real-world scans can be further degraded by noises and occlusions, which complicate purely geometry-based methods [78]. Last but not least, most existing datasets lack paired RGB images with annotated 3D affordance regions tied to interaction cues, precluding learning models from end-to-end training and evaluation [76]. Without a unified solution to tackle these issues, creating an affordance bridge to localization affordances from visual signifiers is a utopian task. The central question is: "*How can we learn representations that match diverse spatial affordances across different scenes from visual signifiers?*"

To tackle this problem, we introduce a new benchmarking dataset, namely *AffordBridge*, with RGB image–point cloud affordance annotations, and propose a visual-guidance reasoning affordance method, so-called *AffordMatcher*, that explicitly grounds visual signifiers into precise spatial affordances. As shown in Table 1, our large-scale benchmark has $317,844$ high-resolution paired samples of 2D-3D representations in 685 indoor scenes, resulting in $291,637$ volumetric masks of functionally interactive elements among 157 object categories through 61 actionable affordances. Building upon this dual modality, we develop our affordance learning model, *AffordMatcher*, to semantically align keypoints in visual signifiers with those in point cloud instances, which accurately localize and segment interactable regions from both modalities, as shown in Fig. 1. To summarize, our contributions are twofold, as follows:

- **Affordance Dataset:** We introduce *AffordBridge*, a large-scale dataset that annotates high-resolution point clouds and RGB images, alongside language-descriptive actions, for spatial affordance localization.
- **Affordance Learning:** We propose *AffordMatcher* for effectively matching affordance regions in point clouds derived from RGB images.

2. Related Work

Affordance Datasets. Existing affordance datasets primarily focus on annotating functional regions for human-object interactions, often at either the pixel or voxel level. Focusing images, EPIC-Aff [42], IIT-AFF [44], and AGD20k [35] provide annotations with instance segmentation masks. Voxel-level annotated datasets, such as PartNet [39], 3DAffordanceNet [14], and AffordPose [21], offer spatial masks for affordances; yet, they are all limited to distinct objects rather than full scenes. Recent works, PIAD [74], LASO [30], and SeqAfford [76], incorporate implicit interactions through single commands or sequences of objects; meanwhile, Scenefun3D [9] and MIPA [17] extend to 3D scenes with fewer samples and limited affordance diversity,

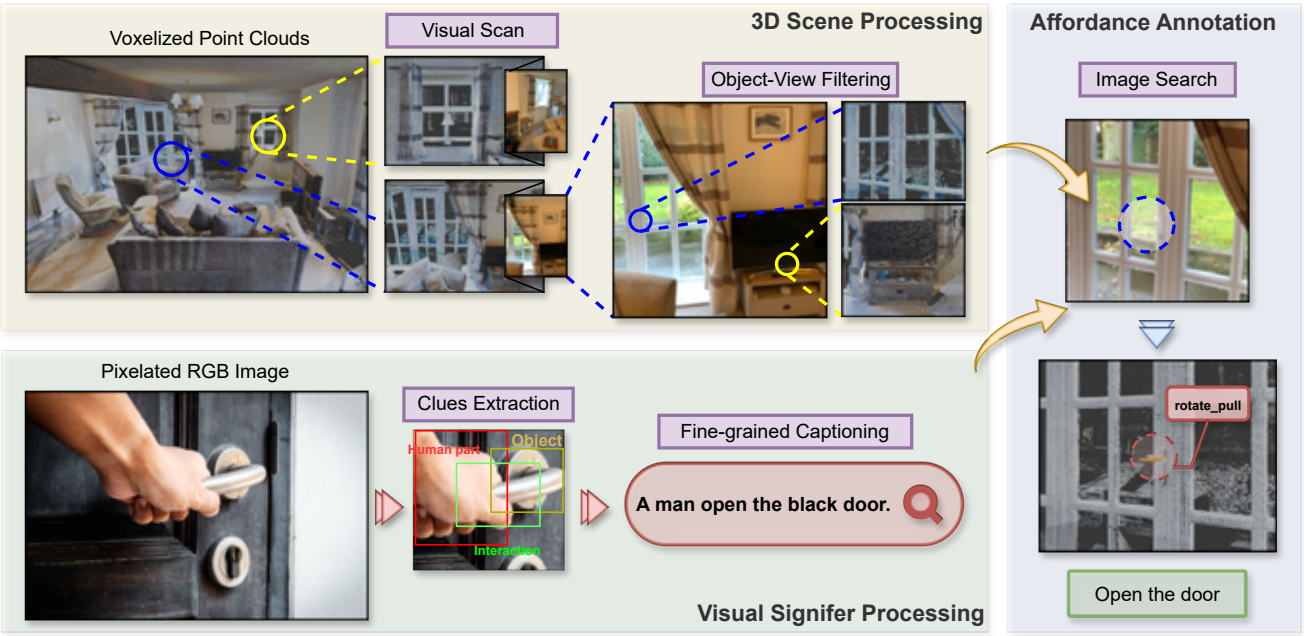


Figure 2. Construction of the *AffordBridge* dataset: Our *AffordBridge* dataset is built through a semi-supervised pipeline linking visual signifiers with 3D affordances. The building process includes (*i*) 3D scene processing via voxelized point clouds with object-view filtering through visual scanning, (*ii*) visual signifiers processing with human-object interaction extraction with fine-grained captioning, and (*iii*) affordance annotation by matching key views to 3D instances for spatial action labeling.

however. Most datasets emphasize object-level affordances (*e.g.*, under $40,000$ samples and fewer than 25 actions) that struggle to capture small functional details in complex scenes and are complicated by the integration of multimodal signifiers. To address these limitations, we propose a large-scale dataset of 3D scenes that incorporates both implicit and explicit descriptions of interactions among instance categories and actions for affordance learning from visual signifiers.

Affordance Learning. Prior studies on affordance learning focused on detecting affordance regions in images [4, 14, 63], while recent research expands affordance learning into the spatial domain [5, 48]. 3D AffordanceNet [10] introduced the first benchmark dataset for learning affordances from 3D objects. Additionally, several methods incorporate additional information, such as images [74], drone-related information [68], or natural language instructions [16, 30], to enhance the reasoning behind affordance regions. However, existing methods are mostly limited to point-level or object-level detection. Recent works have explored 3D indoor scene understanding guided by open-vocabulary [12, 46, 57, 62]. Recently, SceneFun3D [9] was proposed as a high-quality dataset for affordance understanding, featuring diverse natural language descriptions. Nevertheless, SceneFun3D [9] only detects the affordances from the text prompt input. In this work, we focus on localizing spatial affordance from visual signifiers that provide human-object interactions.

Semantic Correspondence. Semantic matching focuses on identifying corresponding elements between the same instances in multiview settings. Traditional approaches achieve this by matching pixels or patches in given image pairs using fine-grained extracted features, which are thus matched through a correlation map using convolutional layers [13, 28, 38, 55], transformer networks [22, 58, 60, 61], or another dedicated backbone [29, 64] between views. Expanding beyond RGB images, modern approaches explore spatial feature matching. For example, DenseMatcher [79] combines the generalization capability of 2D foundation models with 3D geometric understanding to match textured 3D object pairs. 2D3D-MATR [27] first establishes coarse correspondences between local patches in the point cloud and image view, then applies multi-scale patch matching to learn global contextual constraints. Here, our approach employs match-to-match attention to analyze cross-modal point cloud-image correspondences through affordances.

3. The *AffordBridge* Dataset

In Fig. 2, we outline our three-stage annotation flow, which includes 3D scene processing (Sec. 3.1), visual signifier processing (Sec. 3.2), and affordance annotation (Sec. 3.3).

Let the colored scene point cloud be $\mathcal{P} = \{(p_i, f_i)\}_{i=1}^{N}$, where $p_i \in \mathbb{R}^3$ denotes the 3D coordinates, and $f_i \in \mathbb{R}^6$ represents per-point features, including RGB colors and surface normals. From raw scans in [9], the point clouds are downsampled to $100,000$ points via voxelization with a voxel size of 0.05 meters [37], while preserving the details of scene representations and maintaining frugality for the affordance learning model. Thus, instance segmentation masks are applied to extract object regions within the scene [24].

3.1. 3D Scene Processing

Visual Scan. Each colored scene point cloud is temporally aligned with the corresponding RGB video sequence $\mathcal{V} = \{v_k\}_{k=1}^{K}$, where each frame v_k represents a visual observation captured from a calibrated camera pose $[R_k \mid t_k]$ using the camera's intrinsic matrix, denoted as K_c. Through SLAM-based trajectory estimation [80], each frame v_k is projected onto its associated 3D segment through depth alignment, ensuring geometric and temporal consistency along the trajectory. Multiview checks are used to remove occlusions and misalignments, yielding clean 2D-3D correspondences, denoted as (P_k, v_k), for downstream annotation.

Object-View Filtering. Each visual scan may contain multiple objects with potential affordances. We detect candidate objects using MobileNet [20] and rank them by spatial location, scale, and contextual relevance. Each detected instance $v_k^{(l)}$ is therefore aligned with its 3D counterpart $P_k^{(l)}$, and inconsistent matches of approximately 15% are manually discarded. To maintain reliability, we ensure inter-annotator agreements with a Kappa score higher than 0.75 [7] among three human experts reviewing the annotations.

3.2. Visual Signfier Processing

Annotation of Visual Signifiers. We adopt interaction images from PIAD [74], retaining only those that depict direct human-object contact. Each image is annotated with three bounding boxes $b = (b_H, b_O, b_I)$ for human, object, and interaction regions, following the notation of MUREN [23]. We refine the interaction box b_I by inspecting model-predicted class scores to precisely localize the contact region. To further address potential ambiguities in visual signifiers, as static images may miss dynamics, we incorporate human pose estimation via OpenPose [3] to annotate keypoint-based hand-object contacts, enhancing the bounding boxes for humans, objects, and contact regions.

Fine-grained Captioning. Next, we generate fine-grained captions using visual signifiers as inputs. Specifically, using the Object Relation Transformer (ORT) [19], we fuse the three bounding boxes b to produce coherent, templated descriptions. As shown in Fig. 2, the caption "*A man opens the black door*" describes the RGB image input. All captions are manually verified for semantic and spatial accuracy.

3.3. Affordance Annotation

To associate textual descriptions with corresponding scene views, we align image-text embeddings using CLIP encoders [52] and retrieve the most relevant key-view by maximizing cosine similarity. The embedding space is refined through contrastive learning [50] to achieve higher similarity between positive image-text pairs and vice versa.

After filtering, each key-view I_i is paired with an affordance action a_i and its corresponding region within $\mathcal{P}$. We use a web-based annotation interface [8], where annotators are able to identify the 3D instance segmentation mask M_i that corresponds to the 2D view with the affordance mask A_i for the affordance action a_i. The mapping between 2D affordances and 3D instances is $M_i = \arg\max_{M_j \in \mathcal{P}} \phi_{\text{sim}}(I_i, M_j)$, where $\phi_{\text{sim}}(I_i, M_j)$ measures the visual-geometric similarity between the key-view and the 3D object. To avoid one-to-many ambiguities, where a single visual signifier may correspond to multiple 3D instances, we allow multi-instance projection by retaining top-3 matches based on CLIP similarity scores. Subsequently, annotators verify and refine these candidates to obtain the most accurate correspondence. The resulting annotations yield object-level affordance masks embedded within full-scene geometry, with approximately 5% of samples re-annotated to resolve ambiguous cases.

Criteria \ Split	Train	Validate	Test	Total
Visual Signifiers	6,416	1,974	1,480	9,870
3D Scenes	448	138	103	689
Affordance Areas	189,564	58,327	43,746	291,637
Total Samples	206.6K	63.6K	47.7K	317.8K

Table 2. Train/validation/test split of the *AffordBridge* dataset.

3.4. Dataset Statistics

Our *AffordBridge* dataset is organized into training, validation, and test sets across three modalities: visual signifiers, 3D scenes, and affordance areas, as shown in Table 2. The visual signifiers subset contains $9,870$ samples, while the 3D scenes subset includes 689 samples. Affordance areas form the largest component, with $291,637$ samples. In total, the dataset comprises 317.8K samples, with 206.6K for training, 63.6K for validation, and 47.7K for testing, providing a sufficient scale for learning spatial affordance.

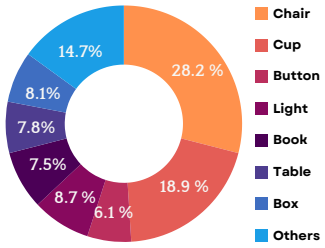


Figure 3. Dataset statistics: Statistics of objects in human-object interactions yielding affordances in our *AffordBridge* dataset.

Fig. 3 illustrates the object distribution across 689 indoor 3D scenes. More specifically, chairs account for 28.2%, cups for 18.9%, and buttons for 8.7%, representing the most frequently interacted objects in the dataset. Lights contribute 7.5%, books 7.8%, tables 8.1%, and boxes 6.1%, each providing essential diversity for modeling functional affordances. The "*Others*" category covers 14.7% of the dataset. In Fig. 3, the object distribution demonstrates the diversity and balance of our *AffordBridge* dataset, supporting both object-centric and scene-level affordance learning.

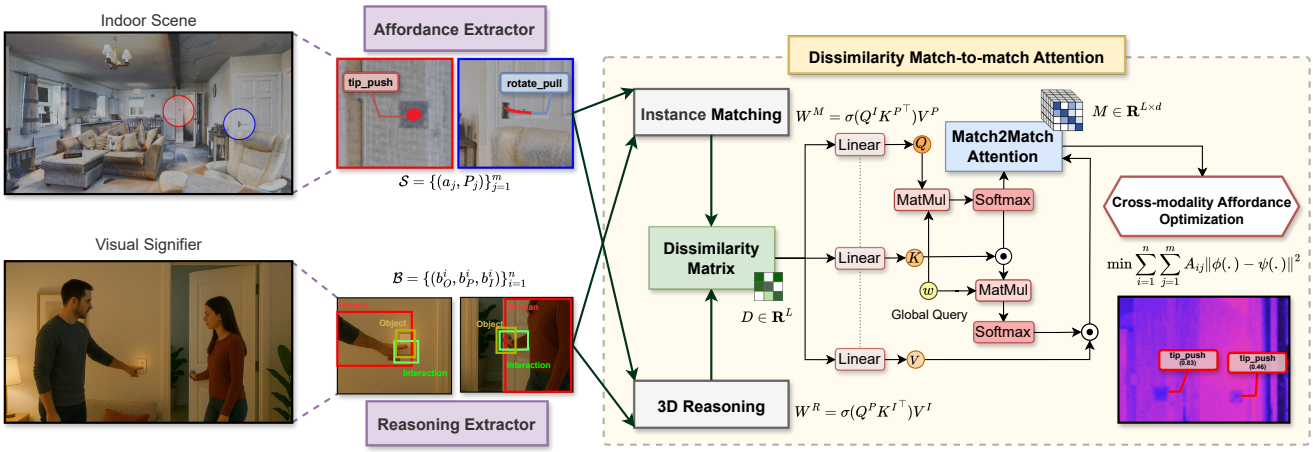


Figure 4. Design architecture of *AffordMatcher*: Given a high-resolution voxelized scene point cloud and a visual signifier, *AffordMatcher* reasons over these inputs for zero-shot affordance segmentation. The *affordance extractor* identifies 3D interactable regions, while the *reasoning extractor* encodes 2D human-object cues. Cross-modal alignment is achieved via instance matching through a dissimilarity matrix. The features from the dissimilarity matrix are thus optimized through match-to-match attention, followed by a zero-shot affordance optimization to localize actionable spatial regions that align with the given signifier.

4. *AffordMatcher*: Affordance Learning in 3D Scenes from Visual Signifiers

We formulate affordance grounding as an alignment problem between visual signifiers and 3D affordance regions. Our objective is to minimize the cross-modal feature discrepancy:

$$\min_{\phi,\psi} \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \left|\phi(b_i) - \psi(a_j, P_j)\right|^2, \quad \text{s. t.} \sum_{j=1}^{m} A_{ij} = 1, \quad \forall i \in \{1, \ldots, n\}, \tag{1}$$

where $\mathcal{P} = \{P_j\}_{j=1}^{m}$ denotes the point cloud composed of local regions P_j, and I is the corresponding RGB image containing n interaction cues $\{b_i\}_{i=1}^{n}$. In Eq. 1, the reasoning extractor $\phi : \mathbb{R}^{d_b} \to \mathbb{R}^d$ encodes each b_i that represents human-part, object, and interaction features. Meanwhile, the affordance extractor $\psi : \mathbb{R}^{d_a} \times \mathbb{R}^3 \to \mathbb{R}^d$ is mapping each 3D region P_j and its action descriptor a_j into the same embedding space. The weight A_{ij} measures the confidence that b_i aligns with the affordance instance (a_j, P_j).

4.1. Instance Matching & 3D Reasoning

Let $F_P \in \mathbb{R}^{m \times N_P}$ and $F_I \in \mathbb{R}^{n \times N_I}$ denote the feature representations of 3D candidate regions $\{M_j\}_{j=1}^{m}$ and visual signifiers $\{b_i\}_{i=1}^{n}$, where N_P and N_I represent the number of point features and the dimensionality of each visual signifier, respectively. We first project them into a shared space of dimension N_D to obtain queries, keys, and values as:

$$Q^{(I)} = F_I W_q^{(I)},\ K^{(P)} = F_P W_k^{(P)},\ V^{(P)} = F_P W_v^{(P)},$$
$$Q^{(P)} = F_P W_q^{(P)},\ K^{(I)} = F_I W_k^{(I)},\ V^{(I)} = F_I W_v^{(I)}. \tag{2}$$

Based on Eq. 2, cross-modal attention is then applied bidirectionally to align 2D and 3D representations as follows:

$$W^{(M)} = \texttt{softmax}\left(Q^{(I)} {K^{(P)}}^\top\right) V^{(P)}, \tag{3a}$$
$$W^{(R)} = \texttt{softmax}\left(Q^{(P)} {K^{(I)}}^\top\right) V^{(I)}, \tag{3b}$$

where $W^{(M)} \in \mathbb{R}^{n \times N_D}$ localizes spatial keypoint features guided by visual signifiers, and $W^{(R)} \in \mathbb{R}^{m \times N_D}$ captures reasoning feedback propagated from the 3D context.

4.2. Dissimilarity Quantification

From Eq. 3, we compute the dissimilarity matrix $D \in \mathbb{R}^{n \times m}$ to quantify the cross-modal correspondence between visual and spatial features. Each entry $D_{ij} \in [0, 1]$ measures the cosine dissimilarity between the i-th and the j-th spatial features:

$$D_{ij} = 1 - \max\left\{0, \frac{W_i^{(M)} \cdot W_j^{(R)}}{\|W_i^{(M)}\|_2 \, \|W_j^{(R)}\|_2}\right\}, \tag{4}$$

where $\cdot$ is the inner product and $\|\cdot\|_2$ is the Euclidean norm.

The dissimilarity matrix D, with each entry as in Eq. 4, is flattened into a length $L = nm$ and projected into an embedding of dimension N_X, yielding $X = DW_X \in \mathbb{R}^{L \times N_X}$. We then apply additive FastFormer-style self-attention [70] over (Q, K, V), defined to be (XW_q, XW_k, XW_v), as:

$$Z = K \odot \left(\sigma(Qw_q)^\top Q\right), \ M = V \odot \left(\sigma(Zw_k)^\top Z\right), \tag{5}$$

where $\sigma(\cdot)$ denotes the element-wise sigmoid, $\odot$ represents the Hadamard product, and $w_q, w_k \in \mathbb{R}^{N_X}$ are learnable parameter vectors. Thus, a multi-head projection produces the final match matrix. With M from Eq. 5, we obtain:

$$\mathcal{M} = \texttt{MultiHead}(M)\, W_h + b_h, \tag{6}$$

where $\mathcal{M}$ is the `Match2Match` attention map, which is processed by the bounding-box and mask prediction heads. To address one-to-many correspondences as described, we apply a soft-thresholding mask on D. High-similarity pairs, $D_{ij} < 0.2$, are allowed to propagate multiple times through $\mathcal{M}$ to seek robust and stable matches in cluttered scenes.

4.3. Cross-modality Affordance Learning

To enable cross-modality affordance learning that solves Eq. 1, we align (*i*) embedding normalization for global consistency (Eq. 7), followed by (*ii*) semantic and geometric embeddings (Eq. 8), (*iii*) bidirectional mapping between modalities (Eq. 9), and (*iv*) cross-modal attention dissimilarity (Eq. 10).

First, let $\phi_i : b_i \mapsto \mathbb{R}^d$ and $\psi_j : (a_j, P_j) \mapsto \mathbb{R}^d$ denote the projection heads for visual signifiers and spatial regions, respectively. Both embeddings are constrained to lie on the unit hypersphere, $\|\phi(b_i)\|_2 = 1$ and $\|\psi(a_j, P_j)\|_2 = 1$, to maintain feature and geometric consistency. The weighted sum of a normalization term and a regularization term results in the embedding regularization loss $\mathcal{L}_{\text{embed}}$:

$$\alpha \left[\sum_{i=1}^{n} (\|\phi_i\|_2 - 1)^2 + \sum_{j=1}^{m} (\|\psi_j\|_2 - 1)^2 \right] + \beta \sum_{\theta} \|\theta\|_F^2, \tag{7}$$

where $\theta \in \Theta_\phi \cup \Theta_\psi$, Θ_ϕ and Θ_ψ represent the trainable parameters of ϕ and ψ, respectively. Then, let M_{ij} denote the FastFormer attention output (Eq. 5) and T_{ij} its pseudo-target from S-CLIP [40], computed as a convex combination of CLIP text embeddings. The alignment loss $\mathcal{L}_{\text{align}}$ is:

$$\mathcal{L}_{\text{align}} = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} \, \|M_{ij} - T_{ij}\|_2^2. \tag{8}$$

Next, with two linear projection heads, $g_{\text{ins}}, g_{\text{r}} : \mathbb{R}^d \to \mathbb{R}^d$, we further enforce bidirectional consistency $\mathcal{L}_{\text{bidir}}$:

$$\mathcal{L}_{\text{bidir}} = \sum_{i,j} A_{ij} \left(\|g_{\text{ins}}(\phi_i) - \psi_j\|_2^2 + \|g_{\text{r}}(\psi_j) - \phi_i\|_2^2 \right) \tag{9}$$

Lastly, we penalize the ReLU-clipped cosine dissimilarity between $W^{(M)}$ and $W^{(R)}$, yielding the dissimilarity loss that explicitly reduces the cross-modal attention $\mathcal{L}_{\text{dissim}}$:

$$\mathcal{L}_{\text{dissim}} = \sum_{i,j} A_{ij} \left[1 - \frac{W_i^{(M)} \cdot W_j^{(R)}}{\|W_i^{(M)}\|_2 \|W_j^{(R)}\|_2} \right]. \tag{10}$$

Overall, the training objective combines all losses with the empirically chosen set of weights $\{\alpha, \beta, \lambda, \gamma, \eta\}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{embed}} + \lambda \, \mathcal{L}_{\text{align}} + \gamma \mathcal{L}_{\text{bidir}} + \eta \, \mathcal{L}_{\text{dissim}}. \tag{11}$$

To train Eq. 11, we employ the reasoning extractor ϕ and the affordance extractor ψ with ViT-B/16 [15] and PointNet++ [51], respectively, each followed by two-layer MLP projection heads, with pose augmentation applied to ϕ.

5. Experiments & Evaluations

5.1. Experiment Setup & Baselines

Implementation Details. In our experiments, the *AffordMatcher* model is trained with RGB-point cloud and description text inputs for 100 epochs on an NVIDIA RTX 3090 GPU with a batch size of 16 and an initial learning rate of 10^{-4} decayed by 0.5 every 30 epochs. The RGB images are resized to 224×224 with visual augmentation, while 3D scenes are voxelized into a 64^3 grid and segmented by a pre-trained 3D model to obtain binary-mask affordance candidates. Our evaluation follows the standardized zero-shot affordance segmentation metrics: mAP@0.25, mAP@0.50, and mAP averaged over IoU thresholds from 0.50 to 0.95.
Baselines. We benchmark *AffordMatcher* against state-of-the-art baselines in functional adaptations of 3D instance segmentation methods, including Mask3D-F, SoftGroup-F, and OpenMask3D-F, as listed in [9], and full pipelines, including AffordPose-DGCNN [21], 3DAffordanceNet [10], PIAD [74], LASO [30], and Ego-SAG [33].

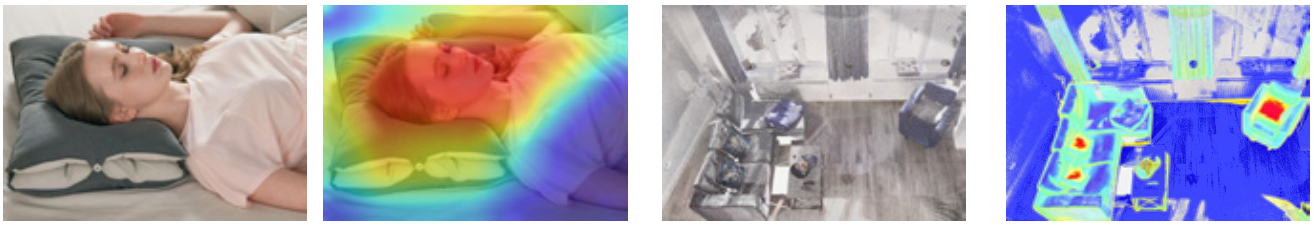
5.2. Quantitative Results

Table 3 reports the performance of our method against state-of-the-art methods on functionality affordance segmentation. Specifically, *AffordMatcher* achieves an overall mAP of 53.4, outperforming the second-best baseline by 7.8 while exhibiting superior localization of functionally relevant regions among both low and high IoU thresholds. The improvement demonstrates the effectiveness of visually guided reasoning in improving spatial affordance localization in 3D scenes, given visual signifiers, under zero-shot settings.

Method \ Metric	mAP	mAP @0.25	mAP @0.50	No. Params	Inference Speed (ms / sample)
Mask3D-F [9]	41.2	58.6	47.1	19.0M	<u>126.2</u>
SoftGroup-F [9]	43.9	60.8	49.3	30.4M	288.0
OpenMask3D-F [9]	<u>45.6</u>	<u>62.1</u>	<u>51.0</u>	39.7M	315.1
APose-DGCNN [21]	29.7	47.6	34.8	**12.5M**	140.2
3DAffordanceNet [14]	34.2	51.3	39.6	<u>15.0M</u>	180.4
PIAD [74]	26.1	44.7	30.5	23.0M	160.9
LASO [30]	37.5	54.2	42.6	21.4M	130.4
Ego-SAG [33]	40.3	56.7	45.1	24.8M	175.3
***AffordMatcher* (Ours)**	**53.4**	**69.7**	**59.5**	20.7M	**112.5**

Table 3. Quantitative results: Performance comparisons of *AffordMatcher* and state-of-the-art methods [9, 14, 21, 30, 33, 74] in terms of mAP, mAP@0.25, mAP@0.50, number of parameters (in millions), and inference speed (in milliseconds per sample).

Also shown in Table 3, *AffordMatcher* attains high accuracy while maintaining computational efficiency. The model contains 20.7 million parameters; for example, fewer than OpenMask3D-F, while achieving faster inference at 112.5 milliseconds per sample. The balanced trade-off between accuracy and efficiency illustrates the scalability of *AffordMatcher* for large-scale and near real-time 3D scene affordance localization.



(a) Visual signifer **(b)** Attention over visual signifer **(c)** High-resolution voxelized scene **(d)** Attention over the given scene

Figure 5. Attention visualization: From the visual signifier in the RGB image and the text "*Rest on Pillow*", *AffordMatcher* focuses on the pillow area in the RGB image and correctly localizes the corresponding affordance regions in the high-resolution voxelized indoor scene.

Variant of Inputs \ Metric	mAP	mAP @0.25	mAP @0.50
pruning RGB image inputs	37.3	52.7	42.1
inpainting human-object interactions	40.9	56.2	45.3
without point cloud downsampling	48.7	65.1	54.2
fine-tuning with PIAD objects	45.3	61.8	50.6
***AffordMatcher* (Ours)**	**53.4**	**69.7**	**59.5**

Table 4. Ablation on different input modalities: Removing visual inputs severely degrades performance, followed by inpainting human-object interactions, fine-tuning with PIAD objects, and thus using raw point cloud. The full *AffordBridge* achieves the highest accuracy through integrated 2D-3D reasoning.

(a) without visual reasoning **(b)** with visual reasoning

Figure 6. t-SNE visualization: Visual reasoning produces more compact and well-separated clusters among different affordance types compared to without visual reasoning.

5.3. Ablation Study

Input Guidance Analysis. As shown in Table 4, we conduct experiments to assess whether the observed performance gains originate from interaction reasoning rather than object recognition. We found that eliminating the 2D branch leads to a significant decrease in mAP to 37.3, indicating the critical role of visual cues. Meanwhile, removing humans and hands from RGB images via inpainting [65] lowers the mAP to 40.9, confirming that action semantics contribute to interaction reasoning. Using raw point clouds of more than $500,000$ points reduces the mAP to 48.7 due to memory constraints. Fine-tuning on the object-centric PIAD dataset produces 45.3 of mAP, validating the advantage of modeling scene-level affordance cues over isolated object interactions.

$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{dissim}}$	$\mathcal{L}_{\text{embed}}$	$\mathcal{L}_{\text{bidir}}$	mAP	mAP @0.25	mAP @0.50
				37.3	52.7	42.1
✓				40.9	56.2	45.3
✓	✓			44.1	60.0	48.7
✓	✓	✓		47.8	63.5	53.0
✓	✓	✓	✓	**53.4**	**69.7**	**59.5**

Table 5. Ablation on loss components: Gradual inclusion of each loss enhances performance, with the full objective achieving the best results. The first row indicates the baseline, which corresponds to the semantic affordance objective [9].

Loss Component Analysis. Table 5 reports the contribution of each loss component. Beginning with a standard semantic baseline, the inclusion of $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{dissim}}$ yields significant improvements, showcasing the importance of cross-instance correspondence modeling. Adding the bidirectional and regularization losses further enhances performance, resulting in a cumulative mAP gain of 16.1, which shows that the complete objective effectively learns structured and discriminative affordance representations.

Attention Visualization. Fig. 5 illustrates how the learned attention maps transfer reasoning cues across modalities. For the "*Rest on Pillow*" example, the visual signifier concentrates on the pillow region. Through this, its spatial attention is able to emphasize the corresponding voxels on seating surfaces, confirming consistent cross-modal learning for human-object affordance localization.

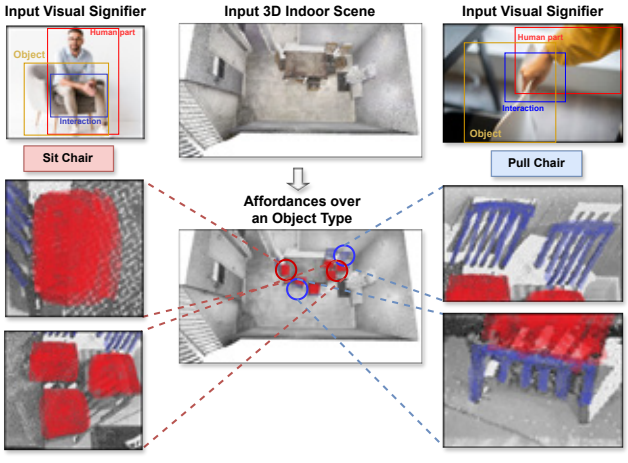


Figure 7. Visualization of distinct affordances: "*Sit*" and "*Pull*" cues on the same chair activate different 3D regions, showcasing that *AffordMatcher* adapts attention to interaction semantics.

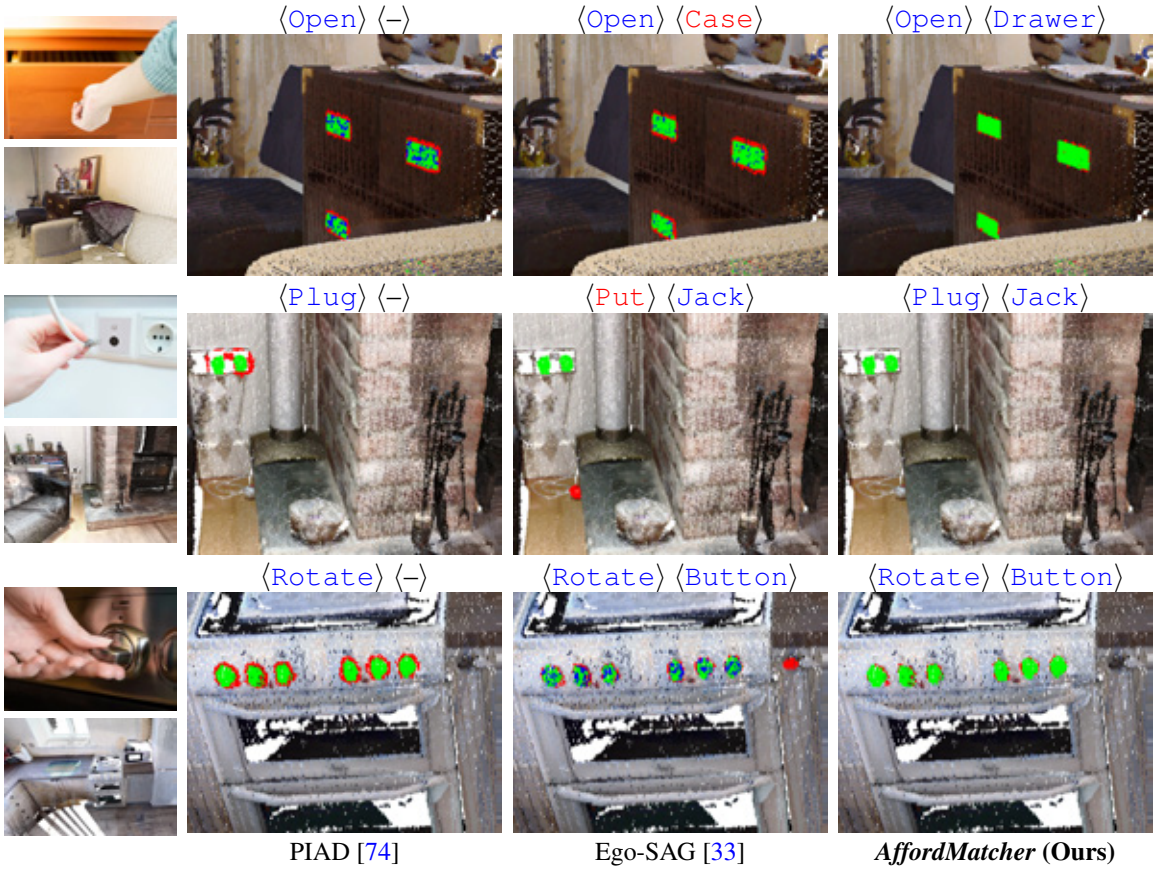


Figure 8. Qualitative results: Our affordance-mask prediction, compared with other baselines. The first column illustrates the inputs, including visual signifiers and 3D scenes. The remaining columns show the affordance segmentation results, including extracted actions over different methods, where **blue texts** indicate correct affordance actions and **red texts** represent wrong ones. In the point clouds, **green areas** denote correct affordance localization, **red** and **blue areas** denote false positives and false negatives, respectively.

Reasoning Analysis. The t-SNE embeddings [66] in Fig. 6 reveal that the reasoning module on Sec. 4.1 produces compact and well-separated clusters, indicating improved affordance discriminability. Meanwhile, Fig. 7 further demonstrates reasoning adaptability by visualizing distinct interaction cues, such as "*Sit*" and "*Pull*", applied to the same object. For the "*Sit*" action, attention focuses on the seat cushion and the frontal area of the chair; whereas for the "*Pull*" action, the attention then shifts toward the upper back and armrest regions, demonstrating that *AffordMatcher* can dynamically adjust its focus according to given visual signifiers while maintaining a consistent geometric understanding.

Qualitative Results. Fig. 8 presents qualitative comparisons between our *AffordMatcher* and PIAD [74] together with Ego-SAG [33]. PIAD tends to under-segment affordance regions, often missing fine interaction details, while Ego-SAG often over-segments them, producing overly broad or redundant affordance masks that lack spatial precision. Notably, *AffordMatcher* is able to generate compact and accurate affordance masks that suit both coarse and fine-grained interaction parts, such as knobs and prongs, while achieving higher spatial precision.

Limitations. Although *AffordMatcher* demonstrates strong cross-modality affordance learning capability, it faces challenges with memory and scalability in highly detailed scenes, resulting in increased computational costs. Some errors also occur, such as overlapping affordances or unclear actions (please see details in our Supplementary Material), which reveal limits in disambiguation and spatial reasoning.

6. Conclusions

In this work, we introduce *AffordMatcher*, an affordance learning method for spatial affordance localization in high-resolution voxelized indoor scenes from visual signifiers through sophisticated cross-modal reasoning. By leveraging the *AffordBridge* dataset and a dissimilarity-based match-to-match attention mechanism, *AffordMatcher* achieves robust zero-shot affordance segmentation across diverse scenes. Experimental results demonstrate consistent gains over state-of-the-art methods in both accuracy and efficiency, validating the effectiveness of reasoning-guided affordance learning. We reserve the task of extending *AffordMatcher* to temporal and interactive scenarios, enabling dynamic affordance reasoning in real-world robotic systems and embodiments.

References

[1] Shikhar Bahl, Russell Mendonca, Lili Chen, et al. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023. 1

[2] Raphaël Braud, Alexandros Giagkos, Patricia Shaw, et al. Robot multimodal object perception and recognition: Synthetic maturation of sensorimotor learning in embodied systems. *IEEE Transactions on Cognitive and Developmental Systems*, 2020. 1

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, et al. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 4

[4] Joya Chen, Difei Gao, Kevin Qinghong Lin, et al. Affordance grounding from demonstration video to target image. In *CVPR*, 2023. 3

[5] Shizhe Chen, Ricardo Garcia, Ivan Laptev, et al. Sugar: Pre-training 3d visual representations for robotics. In *CVPR*, 2024. 3

[6] Nhat Chung, Taisei Hanyu, Toan Nguyen, Huy Le, et al. Rethinking progression of memory state in robotic manipulation: An object-centric perspective. In *AAAI*, 2026. 1

[7] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1960. 4

[8] Angela Dai, Angel X Chang, Manolis Savva, et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 4

[9] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, et al. Scenefun3d: fine-grained functionality and affordance understanding in 3d scenes. In *CVPR*, 2024. 2, 3, 6, 7

[10] Shengheng Deng, Xun Xu, Chaozheng Wu, et al. 3d affordancenet: A benchmark for visual object affordance understanding. In *CVPR*, 2021. 2, 3, 6

[11] Tharindu Dharmasiri et al. Cross-modal self-training: Aligning images and pointclouds to learn classification without labels. In *CVPRW*, 2024. 2

[12] Runyu Ding, Jihan Yang, Chuhui Xue, et al. Pla: Language-driven open-vocabulary 3d scene understanding. In *CVPR*, 2023. 3

[13] Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Fine-grained visual classification using self assessment classifier. In *IEEE Conference on Artificial Intelligence (CAI)*, 2024. 3

[14] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018. 2, 3, 6

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*, 2020. 6

[16] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, et al. OpenNeRF: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Rendered Novel Views. In *ICLR*, 2024. 3

[17] Xianqiang Gao, Pingrui Zhang, Delin Qu, et al. Learning 2d invariant affordance knowledge for 3d affordance grounding. In *AAAI*, 2025. 2

[18] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Houghton Mifflin, 1979. 1

[19] Simao Herdade, Armin Kappeler, Kofi Boakye, et al. Image captioning: Transforming objects into words. *NIPS*, 2019. 4

[20] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, 2017. 4

[21] Juntao Jian, Xiuping Liu, Manyi Li, et al. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *ICCV*, 2023. 2, 6

[22] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *CVPR*, 2022. 3

[23] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *CVPR*, 2023. 4

[24] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, et al. Oneformer3d: One transformer for unified point cloud segmentation. In *CVPR*, 2024. 3, 1

[25] Olivia Y Lee, Annie Xie, Kuan Fang, et al. Affordance-guided reinforcement learning via visual prompting. *arXiv*, 2024. 1

[26] Gen Li, Nikolaos Tsagkas, Jifei Song, et al. Learning precise affordances from egocentric videos for robotic manipulation. In *ICCV*, 2025. 2

[27] Minhao Li, Zheng Qin, Zhirui Gao, et al. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *ICCV*, 2023. 3, 2

[28] Shuda Li, Kai Han, Theo W Costain, et al. Correspondence networks with adaptive neighbourhood consensus. In *CVPR*, 2020. 3

[29] Xinghui Li, Jingyi Lu, Kai Han, et al. Sd4match: Learning to prompt stable diffusion model for semantic matching. In *CVPR*, 2024. 3

[30] Yicong Li, Na Zhao, Junbin Xiao, et al. Laso: Language-guided affordance segmentation on 3d object. In *CVPR*, 2024. 2, 3, 6

[31] Yong-Lu Li, Xinpeng Liu, Han Lu, et al. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2

[32] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, et al. Multi-modal situated reasoning in 3d scenes. *NIPS*, 2024. 2

[33] Cuiyu Liu, Wei Zhai, Yuhang Yang, et al. Grounding 3d scene affordance from egocentric interactions. *arXiv*, 2024. 6, 8, 2

[34] Timo Lueddecke, Tomas Kulvicius, and Florentin Woergoetter. Context-based affordance segmentation from 2d images for robot actions. *Robotics and Autonomous Systems*, 2019. 2

[35] Hongchen Luo, Wei Zhai, Jing Zhang, et al. Learning affordance grounding from exocentric images. In *CVPR*, 2022. 2

[36] Hongchen Luo, Wei Zhai, Jing Zhang, et al. Leverage interactive affinity for affordance learning. In *CVPR*, 2023. 2

[37] Aihua Mao, Biao Yan, Zijing Ma, et al. Denoising point clouds in latent space via graph convolution and invertible neural network. In *CVPR*, 2024. 3

[38] Juhong Min, Jongmin Lee, Jean Ponce, et al. Learning to compose hypercolumns for visual correspondence. In *ECCV*, 2020. 3

[39] Kaichun Mo, Shilin Zhu, Angel X Chang, et al. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019. 2

[40] Sangwoo Mo, Minkyu Kim, Kyungmin Lee, et al. S-clip: Semi-supervised vision-language learning using few specialist captions. *NIPS*, 2023. 6

[41] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *NIPS*, 2020. 1

[42] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, et al. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020. 2

[43] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Detecting object affordances with convolutional neural networks. In *IROS*, 2016. 2

[44] 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 *IROS*, 2017. 2

[45] Nghia Nguyen, Minh Nhat Vu, Baoru Huang, An Vuong, Ngan Le, Thieu Vo, and Anh Nguyen. Lightweight language-driven grasp detection using conditional consistency model. In *IROS*, 2024. 2

[46] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, et al. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, 2024. 3

[47] Toan Nguyen, Minh Nhat Vu, An Vuong, et al. Open-vocabulary affordance detection in 3d point clouds. In *IROS*, 2023. 2

[48] Chuanruo Ning, Ruihai Wu, Haoran Lu, et al. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *NIPS*, 2023. 3

[49] A Norman Donald. *The design of everyday things*. MIT Press, 2013. 1

[50] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018. 4

[51] Charles Ruizhongtai Qi, Li Yi, Hao Su, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NIPS*, 2017. 6

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[53] Aditya Raikwar, Newton D'Souza, Ciana Rogers, et al. Cubevr: Digital affordances for architecture undergraduate education using virtual reality. In *VR*, 2019. 1

[54] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCVW*, 2016. 2

[55] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, et al. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 3

[56] Yawen Shao, Wei Zhai, Yuhang Yang, et al. Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. In *CVPR*, 2025. 2

[57] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, et al. Language embedded 3d gaussians for open-vocabulary scene understanding. In *CVPR*, 2024. 3

[58] Jiaming Sun, Zehong Shen, Yuang Wang, et al. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 3

[59] Jiahao Sun, Chunmei Qing, Junpeng Tan, et al. Superpoint transformer for 3d scene instance segmentation. In *AAAI*, 2023. 1

[60] Yixuan Sun, Dongyang Zhao, Zhangyue Yin, et al. Correspondence transformers with asymmetric feature learning and matching flow super-resolution. In *CVPR*, 2023. 3

[61] Yixuan Sun, Zhangyue Yin, et al. Pixel-level semantic correspondence through layout-aware representation learning and multi-scale matching integration. In *CVPR*, 2024. 3

[62] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, et al. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NIPS*, 2023. 3

[63] Jiajin Tang, Ge Zheng, Jingyi Yu, et al. Cotdet: Affordance knowledge prompting for task driven object detection. In *ICCV*, 2023. 3

[64] Luming Tang, Menglin Jia, Qianqian Wang, et al. Emergent correspondence from image diffusion. *NIPS*, 2023. 3

[65] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 2004. 7

[66] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 8

[67] Tuan Van Vo, Minh Nhat Vu, Baoru Huang, Toan Nguyen, Ngan Le, Thieu Vo, and Anh Nguyen. Open-vocabulary affordance detection using knowledge distillation and text-point correlation. In *ICRA*, 2024. 2

[68] Nghia Vu, Tuong Do, Dzung Tran, Binh X Nguyen, Hoan Nguyen, Erman Tjiputra, Quang D Tran, Hai Nguyen, and Anh Nguyen. Aeroscene: Progressive scene synthesis for aerial robotics. In *ICRA*, 2026. 3

[69] An Dinh Vuong, Minh Nhat Vu, Baoru Huang, et al. Language-driven grasp detection. In *CVPR*, 2024. 2

[70] Chuhan Wu, Fangzhao Wu, Tao Qi, et al. Fastformer: Additive attention can be all you need. *arXiv*, 2021. 5

[71] Ruihai Wu, Kai Cheng, Yan Zhao, et al. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *NIPS*, 2023. 1, 2

[72] Ruinian Xu, Fu-Jen Chu, Chao Tang, et al. An affordance keypoint detection network for robot manipulation. *RA-L*, 2021. 1

[73] Kashu Yamazaki, Taisei Hanyu, Khoa Vo, et al. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *ICRA*, 2024. 2

[74] Yuhang Yang, Wei Zhai, Hongchen Luo, et al. Grounding 3d object affordance from 2d interactions in images. In *ICCV*, 2023. 2, 3, 4, 6, 8

[75] Dong Woo Yoo, Sakib Reza, Nicholas Wilson, et al. Augmenting learning with augmented reality: Exploring the affordances of ar in supporting mastery of complex psychomotor tasks. *arXiv*, 2023. 1

[76] Chunlin Yu, Hanqing Wang, Ye Shi, et al. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *CVPR*, 2025. 2

[77] Andy Zeng, Shuran Song, Kuan-Ting Yu, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 2022. 1

[78] Zeyu Zhang, Lexing Zhang, Zaijin Wang, et al. Part-level scene reconstruction affords robot interaction. In *IROS*, 2023. 2

[79] Junzhe Zhu, Yuanchen Ju, Junyi Zhang, et al. Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. *arXiv*, 2024. 3

[80] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *CVPR*, 2022. 4

AffordMatcher: Affordance Learning in 3D Scenes from Visual Signifiers

Supplementary Material

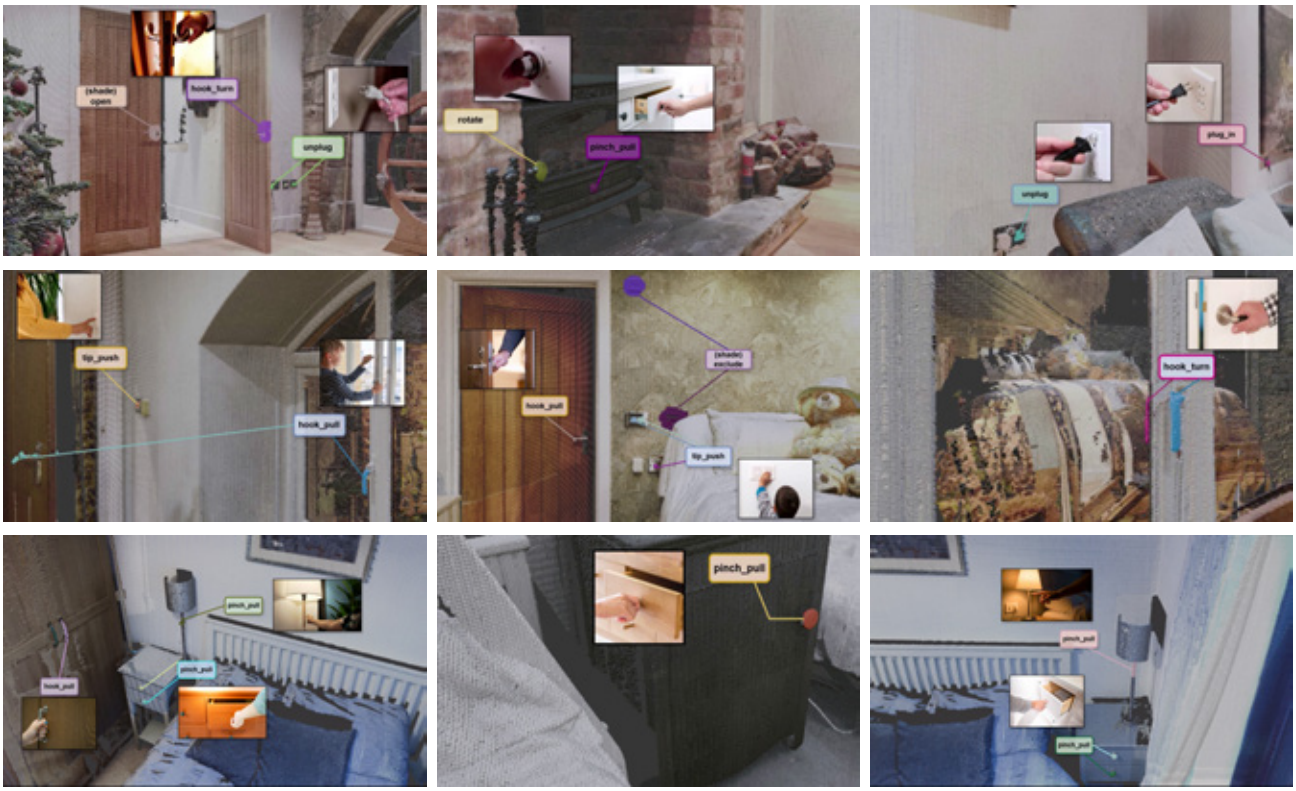


Figure 9. Sample visualization of *AffordBridge* dataset.

Abstract

This supplementary material provides additional dataset analysis, empirical insights, and qualitative evaluations to complement our paper. We begin by providing more details about our AffordBridge dataset, which features visualizations that illustrate affordance areas matched across multimodal inputs, including visual signifiers, textual descriptions, and action labels. We then discuss potential dataset usages, highlighting research directions in 3D scene understanding and human-scene interaction enabled by our annotations. Next, we present results from our user study evaluating the perceptual quality of our proposed AffordMatcher against baselines, demonstrating superior correctness and interpretability. Finally, we visualize representative failure cases of our method to underscore known limitations, such as handling sparse point clouds, complex actions, and ambiguous visuals, thereby reinforcing key observations and identifying opportunities for future work. Please see our video demonstration for a more interactive experience.

7. Dataset Visualization

Fig. 9 provides a sample visualization of the *AffordBridge* dataset, highlighting affordance areas and their corresponding action phrases across different modalities. This figure presents an input scene alongside affordance areas matched to a visual signifier, demonstrating how environmental cues are linked to potential interactions. For more details, please visit our demonstration video.

8. Potential Usages of Our Dataset

Our introduced *AffordBridge* dataset includes annotations for both 3D scenes and RGB images to support affordance reasoning. Below, we outline several exciting research directions that can benefit from leveraging our dataset:

- **3D Scene Understanding.** Traditional approaches to 3D scene analysis often focus on the instance level [24, 59]. By providing annotations for interactive elements on objects, our dataset opens opportunities for addressing various tasks in 3D scene understanding, such as 3D object detection and segmentation.

- **Robotic Manipulation.** Robots with affordance-aware systems can perform tasks more naturally, such as grasping [69], opening [71], or assembling objects in complex, unstructured environments [73]. The release of our dataset could help robotic systems to understand the purpose and function of objects in a 3D context.
- **Human-Scene Interaction.** With affordance masks for 3D indoor scenes and bounding boxes for RGB images, researchers can gain deeper insights into the functional regions of objects and their interactions with humans. This can contribute to the development of more robust human-object interaction models that integrate 2D and 3D data [27, 31], facilitating unified interaction reasoning [32].

9. *AffordMatcher* Analysis

User Study. We conducted a user study to evaluate the perceptual quality of semantic affordance masks produced by our proposed *AffordMatcher* versus three baselines, including Mask3D-F [9], PIAD [74], and Ego-SAG [33]. Twenty experts in 3D vision each reviewed 40 scenes (10 per method), rating the correctness of each affordance mask on a 5-point Likert scale (1 = completely incorrect, 5 = perfect) and selecting the single best segmentation per scene.

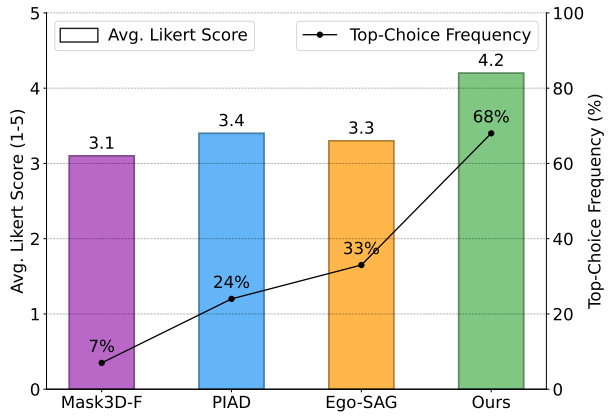


Figure 10. User study results in interaction matching criteria.

Fig. 10 presents the aggregated results: average Likert scores (bars) and the frequency each method was chosen as the top segmentation (line). Our approach achieved an average rating of 4.2, substantially higher than Mask3D-F (3.1), Ego-SAG (3.3), and PIAD (3.4), and was selected as best in 68% of trials, significantly outperforming all baselines ($p < 0.01$, paired t-test).

Participants noted that our masks more accurately captured fine-grained affordance regions, such as chair seats or door handles, and avoided spurious activations common in baseline outputs. This confirms that our *AffordMatcher* not only improves metric performance but also delivers actionable affordance in 3D point clouds.

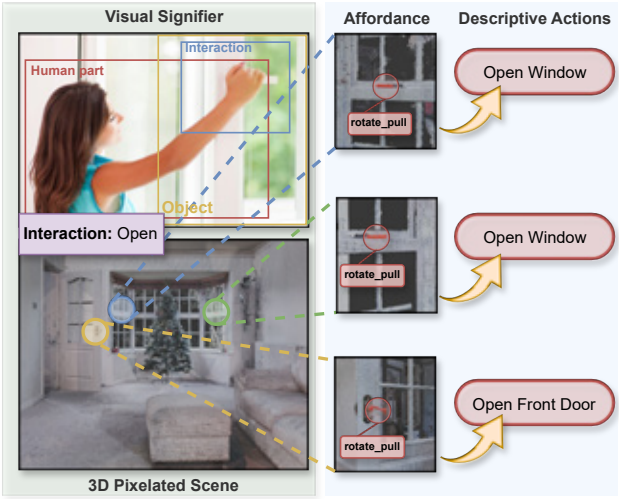


Figure 11. Affordance prediction for a specific interaction, but results in different objects.

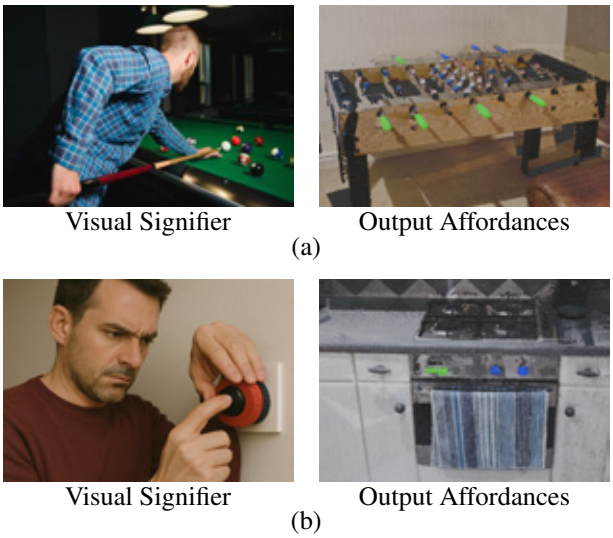


Figure 12. Fail cases of our method: the visual signifier provides a complex action, which causes failure in reasoning. **Green** areas denote correct predictions; **red** and **blue** areas are false positives and false negatives, respectively.

One-to-many Analysis. In Fig. 11, we show the model's ability to localize the support regions required for a single interaction ("Open") across multiple object instances within the same scene. From the top to bottom rows, we feed the network the same 2D visual cue-an outstretched hand poised to open-with the corresponding 3D voxelized indoor environment. The resulting affordance predictions correctly highlight the window latch, the second window's handle, and finally the front door's knob, each delineated by high-response voxels in the 3D scene. These results show that our *AffordMatcher* flexibly generalizes the "pitch_pull" action to match with the interaction, successfully in semantically analogous parts on different objects, even when their appearance, scale, and orientation vary significantly.

Fail Cases. As outlined in the main paper, our method exhibits known limitations related to the challenge of semantic grounding, which are further exemplified through representative failure cases in Fig. 12. Specifically, such failure cases involve nuanced contextual cues or ambiguous object interactions that current models struggle to resolve without task-specific guidance. For example, in the Fig. 12a, the model failed to analyze the “push" action when the man is playing billiards is from which side, or in the Fig. 12b, the model failed to distinguish whether the action in the visual signifier is the “rotate" or the “push". Nonetheless, these examples serve to reinforce the overall generality of our approach under standard conditions, while motivating future work focused on enhancing semantic grounding and model adaptability in more challenging scenarios.