Open-Vocabulary 3D Affordance Understanding via Functional Text Enhancement and Multilevel Representation Alignment

Lin Wu
James Watt School of Engineering, University of Glasgow
Glasgow, United Kingdom
l.wu.1@research.gla.ac.uk

Wei Wei
James Watt School of Engineering, University of Glasgow
Glasgow, United Kingdom
w.wei.1@research.gla.ac.uk

Peizhuo Yu
James Watt School of Engineering, University of Glasgow
Glasgow, United Kingdom
2831853Y@student.gla.ac.uk

Jianglin Lan*
James Watt School of Engineering, University of Glasgow
Glasgow, United Kingdom
Jianglin.Lan@glasgow.ac.uk

Abstract

Understanding 3D affordance is essential for agents to effectively interact with real-world environments, encompassing tasks such as manipulation and navigation. Existing methods typically support open-vocabulary queries through label-based language descriptions but often suffer from limited generalization and weak discriminative ability in their representations. However, affordance understanding requires constructing a coherent semantic landscape from fragmented linguistic expressions—one that preserves intra-class diversity while minimizing inter-class overlap. To address these challenges, we introduce Aff3DFunc, a framework designed to enhance the alignment between affordance and 3D geometry. It begins with a functional text enhancement module grounded in the Information Bottleneck (IB) principle, which strategically enriches affordance semantics by maximizing both relevance and diversity. A dual-encoder architecture is then employed to extract embeddings from both point clouds and text. To bridge the modality gap, we further propose a multilevel representation alignment strategy that incorporates supervised contrastive learning, reinforcing semantic–geometric correspondence in a part-to-whole manner. Extensive experiments demonstrate that our approach significantly enhances the understanding of affordance complexity. The learned representations exhibit high adaptability to diverse text queries, particularly in zero-shot settings. Furthermore, the real-world robot validation confirms that our method improves affordance understanding, enabling more fine-grained manipulation tasks. Project website: https://wulin97.github.io/aff3dfunc.

CCS Concepts

• **Computing methodologies** → *Appearance and texture representations*; *Vision for robotics*; *3D imaging*.

Keywords

Open-vocabulary 3D affordance detection, robotic manipulation

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755239

ACM Reference Format:
Lin Wu, Wei Wei, Peizhuo Yu, Jianglin Lan. 2025. Open-Vocabulary 3D Affordance Understanding via Functional Text Enhancement and Multilevel Representation Alignment. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755239

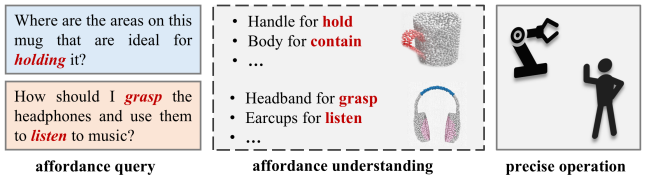


Figure 1: Affordance understanding: Highlight relevant points on a 3D object based on query text, enabling the agent to perform fine-grained and context-specific operations.

1 Introduction

Affordance is a fundamental concept in both perception and robotics, referring to the potential interactions between an agent and its environment [16, 17]. Objects provide opportunities for action based on the agent's capabilities and goals. Given that the real world is inherently three-dimensional, we focus on 3D open-vocabulary affordance understanding—specifically, grounding the regions of point cloud objects to corresponding textual descriptions of affordances, as illustrated in Fig 1. This understanding is crucial for applications in robot manipulation [19, 33, 43] and navigation [30], enhancing scene comprehension and improving the adaptability of autonomous systems in complex environments [3, 10, 48].

Existing methods for 3D affordance understanding predominantly rely on supervised learning that uses large-scale datasets to map the object point cloud to predefined affordance labels. These approaches typically employ Convolutional Neural Networks (CNNs)-based or Transformer-based architectures to extract geometric features from 3D objects [42, 45, 47, 52].With the emergence of foundation models and large language models (LLMs), there is increasing interest in leveraging their semantic capabilities for affordance understanding, with the expectation of improving generalization to unseen scenarios [40]. Recent advances in affordance detection have shown significant progress, exemplified by OpenAD [34],

which integrates the CLIP text encoder [37] with PointNet++ [36] to enhance affordance understanding in point cloud objects. This approach not only provides a more comprehensive grasp of diverse affordances but also demonstrates adaptability to novel affordance labels.

The field of affordance understanding is advancing rapidly, yet substantial challenges remain. Affordances relate to objects as adverbs relate to verbs. [12]: a single object may exhibit multiple affordances, and the same affordance can generalize across diverse objects. This many-to-many relationship poses a fundamental challenge for open-vocabulary detection. While recent work seeks to align multimodal representations, the limited availability of labeled data makes direct learning of affordance semantics difficult. A key step toward robust alignment is to first ***characterize the semantic space of affordances more systematically—beyond raw labels—through principled textual construction***.

To address these challenges, we introduce Functional Text Enhancement (FTE), *an information bottleneck-based strategy that enriches description perspectives and affordance content*. Furthermore, we propose a comprehensive framework that integrates *multilevel representation alignment, supervised contrastive learning, and a cross-attention mechanism* to distinguish similar affordance categories and navigate complex semantic relationships within 3D point clouds in a coarse-to-fine manner. The main contributions of this paper are summarized as follows:

- We propose *Aff3DFunc*, a lightweight open-vocabulary framework for 3D affordance understanding that jointly models affordance semantics and object geometry, enabling generalization to unseen affordances.
- We introduce a Functional Text Enhancement (FTE) strategy based on the information bottleneck principle that systematically characterizes the semantic space by modulating intra-class diversity and inter-class separability.
- We develop a multilevel representation learning scheme that integrates cross-entropy and supervised contrastive losses to establish consistent correspondences between modalities.
- Extensive evaluations, including real-world robotic trials, demonstrate the effectiveness and generalization capability of *Aff3DFunc*, particularly in zero-shot scenarios.

2 Related Work

2.1 Affordance Detection in Point Cloud

Building on the progress in 2D affordance detection [21, 31, 44], the introduction of *3D AffordanceNet*[11] marked a pivotal step toward 3D affordance understanding by mapping affordance labels to point cloud geometry, enabling robots to reason about object utility in 3D space. Subsequent works have expanded this direction by exploring the correspondence between geometric features and affordances through various paradigms. For example, Yang et al.[46] proposed *LEMON* to jointly predict human-object interactions in 3D, while Mur-Labadia et al.[30] introduced a multi-label segmentation framework that grounds affordances from egocentric interaction videos using 3D scene context. Geng et al. [15] proposed *GAParts*, a part-based framework that detects affordances via predefined part categories and strong part supervision. More recently, Nguyen et al.[34] proposed *OpenAD*, an open-vocabulary framework for discovering unseen affordances, and Van et al. [40] enhanced representation alignment via knowledge distillation and text-point correlation. Despite these advances, existing models still struggle in open-vocabulary settings, where affordance categories are often semantically ambiguous or not directly observable from geometry alone. Our method addresses these challenges by explicitly modeling the relationship between affordance semantics and point cloud geometry through *functional text enhancement* and *multi-level representation alignment*.

2.2 Contrastive Representation Learning

Representation learning has been widely explored across supervised, unsupervised, and contrastive learning frameworks [5, 13, 35, 41]. Among these, CLIP [37] and BLIP [22] are two landmark models that advance joint vision-language representation by optimizing contrastive loss on paired image-text data. CLIP, in particular, learns universal cross-modal features that generalize well to downstream tasks without the need for fine-tuning. However, these frameworks operate primarily in self-supervised batch contrastive settings and are less effective in supervised scenarios where label information is available. To address this, Khosla et al. [20] introduced *supervised contrastive learning*, which incorporates label supervision to enhance representation discrimination. In the context of 3D affordance understanding, we adopt similar principles to capture the relationship between affordance semantics and point cloud geometry—pulling together point clusters from the same class while pushing apart those from different classes—to learn discriminative and generalizable features.

2.3 Textual Characterization of Affordance

Common approaches to describing object functionalities include *labels*, *questions*, and *images*, each with inherent limitations. Luo et al. [27] leveraged images for 3D affordance detection by capturing spatial and geometric details, but their method relies heavily on visual affordance cues, limiting generalization. Li et al. [23] introduced questions to enable deeper reasoning, yet such formats increase task complexity and lack directness. Nguyen et al. [34] adopted labels for their simplicity and efficiency, but these are often too coarse to capture functional nuances. Phrases offer a promising alternative, balancing conciseness with semantic richness. Chu et al. [8] explored this direction by embedding affordance labels via large language models (LLMs) to build open-vocabulary affordance representations. However, their approach remains constrained by low efficiency and lacks theoretical grounding for real-world robotic applications. Similarly, Lu et al. [26] proposed a phrase-based modeling framework, but their method suffers from limited sample efficiency and unclear phrase selection criteria. To address these issues, we propose a more robust affordance characterization strategy based on the *information bottleneck* principle, which systematically identifies semantically salient and discriminative descriptions while filtering out redundancy and noise.

3 Problem Description and Methodology

3.1 Problem Description

In 3D open-vocabulary affordance understanding, given an input point cloud $P = \{p_1, p_2, \ldots, p_n\}$ comprising n unordered points,

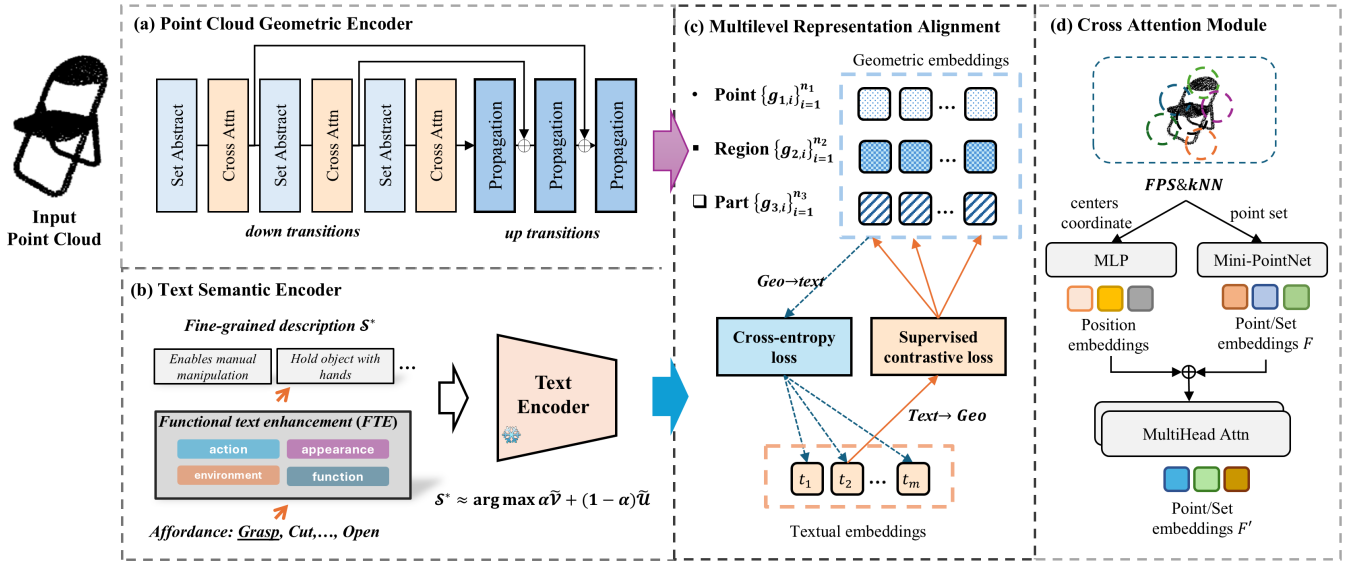


Figure 2: The proposed framework *Aff3DFunc* includes: (a) Point Cloud Encoder, extracting geometric features from input point clouds; (b) Text Encoder, where the FTE module enriches affordance semantics via fine-grained descriptions; (c) Representation Alignment, aligning multimodal embeddings with cross-entropy and supervised contrastive losses across multiple levels; (d) Cross Attention, enhancing geometric features via point-wise relationship modeling using Multi-Head Attention.

where each point $p_i \in \mathbb{R}^3$, $i = 1, \ldots, n$ is defined by its Euclidean coordinates, we aim to discover point-wise affordance through a natural language query $T = \{t_1, t_2, \ldots, t_m\}$. Unlike traditional fixed-label approaches [14, 39], our method allows m to be theoretically unlimited and the query can take any textual form (e.g. label, question, etc.). Our goal is to integrate the affordance description, i.e. *functional text*, with the object point cloud. To achieve this, we use a combination of a point cloud network and a text encoder, which extract geometric and semantic features, respectively. Additionally, we introduce a novel multilevel representation alignment approach and a cross-attention module to connect the comprehensive geometric information with the proposed functional text description. The overall framework of our method is illustrated in Fig. 2.

3.2 Functional Text Enhancement

3.2.1 Mutual Information-guided Description Selection. Text labels provide a simplified representation of affordances but fail to capture their full complexity [7]. To address this limitation, we propose a **Functional Text Enhancement** approach that enriches affordance descriptions by leveraging the **Information Bottleneck (IB) principle** [9, 25]. The IB seeks a compressed representation that retains maximal information about a target variable [18]. In our context, for a given affordance $\mathcal{A}$ with label c, we aim to select a subset of descriptions $\mathcal{S}^* \subseteq \mathcal{S}$ that maximizes the mutual information with $\mathcal{A}$, while minimizing redundancy with the full concept set $\mathcal{S}$. Formally, we define the optimization problem as follows:

$$\mathcal{S}^* = \arg\max_{\mathcal{S}' \subseteq \mathcal{S}} \left[I(\mathcal{A}, \mathcal{S}') - \beta I(\mathcal{S}'; \mathcal{S}) \right], \tag{1}$$

where β is a Lagrange multiplier that balances relevance and redundancy.

However, directly computing $I(\mathcal{A}; \mathcal{S}')$ is intractable due to the high-dimensional and complex nature of the language space. To address this, we introduce an intermediate variable: a set of affordance descriptions $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, where each d_i captures some aspects of the affordance. By controlling the prompting, we generate $\mathcal{D}$ to approximate the true affordance content $\mathcal{A}$. Under the assumptions of conditional independence and convergence, as $n \to \infty$, the mutual information $I(\mathcal{D}; \mathcal{S})$ converges to $I(\mathcal{A}; \mathcal{S})$, i.e., $\lim_{n\to\infty} I(\mathcal{D}; \mathcal{S}) = I(\mathcal{A}; \mathcal{S})$ [38]. With the set $\mathcal{D}$, the IB objective (1) can be reformulated as:

$$\mathcal{S}^* = \arg\max_{\mathcal{S}' \subseteq \mathcal{S}} \left[I(\mathcal{D}, \mathcal{S}') - \beta I(\mathcal{S}'; \mathcal{S}) \right]. \tag{2}$$

To select effective affordance descriptions, we approximate the mutual information using two complementary metrics: intra-class variance ($\mathcal{V}$) and inter-class separability ($\mathcal{U}$).

Maximizing $\mathcal{V}$ encourages a broad semantic coverage within each affordance category, aligning with the goal of increasing the mutual information between descriptions and affordances, $I(\mathcal{D}; \mathcal{S}^*)$. Conversely, maximizing $\mathcal{U}$ enhances the distinctiveness between affordance categories, effectively reducing redundancy and lowering the mutual information between sampled and original description sets, $I(\mathcal{S}^*; \mathcal{S})$. Formally, these relations can be understood as:

$$I(\mathcal{D}; \mathcal{S}^*) \geq H(\mathcal{S}^*) - H(\mathcal{S}^*|\mathcal{D}), I(\mathcal{S}^*; \mathcal{S}) \leq H(\mathcal{S}^*) - H(\mathcal{S}^*|\mathcal{S}), \tag{3}$$

where $H(\cdot)$ denotes entropy. Maximizing $\mathcal{V}$ increases $H(\mathcal{S}^*)$, while maximizing $\mathcal{U}$ decreases the corresponding conditional entropy terms. Therefore, jointly optimizing $\mathcal{V}$ and $\mathcal{U}$ approximates maximizing $I(\mathcal{D}; \mathcal{S}^*) - \beta I(\mathcal{S}^*; \mathcal{S})$, providing an effective and tractable surrogate for mutual information optimization in high-dimensional semantic spaces, with a principled grounding in the IB framework.

3.2.2 Scoring and Sampling Strategy. Several candidate metrics exist for quantifying semantic diversity and class separability. For instance, *pairwise cosine similarity* offers a straightforward estimate of intra/inter-class cohesion, but may be overly sensitive to local variations and less reflective of global structure. Alternatively, *variance-based measures* and *softmax-normalized class separability scores* provide a more stable signal, especially when evaluating pooled embeddings or guiding sample selection. To ensure a balanced trade-off and mitigate scale differences, we normalize both metrics before combining them into a single utility score for pool construction and sampling evaluation:

$$\text{Score}(\mathcal{S}) = \alpha \cdot \tilde{\mathcal{V}}(\mathcal{S}) + (1-\alpha) \cdot \tilde{\mathcal{U}}(\mathcal{S}) \tag{4}$$

where $\tilde{\mathcal{V}}(\mathcal{S})$ and $\tilde{\mathcal{U}}(\mathcal{S})$ are the normalized intra-class variance and inter-class separability, respectively, and α balances their importance. Their normalized values are computed as:

$$\tilde{\mathcal{V}}(\mathcal{S}) = \frac{\mathcal{V}(\mathcal{S}) - \min(\mathcal{V})}{\max(\mathcal{V}) - \min(\mathcal{V})}, \tilde{\mathcal{U}}(\mathcal{S}) = \frac{\mathcal{U}(\mathcal{S}) - \min(\mathcal{U})}{\max(\mathcal{U}) - \min(\mathcal{U})} \tag{5}$$

To mitigate the hallucination effect and inherent uncertainty of LLMs, and building on existing work [26, 38], the proposed FTE focuses on generating descriptions from four core perspectives: *actions*, *functions*, *appearance*, and *environment*. Specifically: *a*) *Actions*: The actions that can be performed on the object, which is associated with the corresponding affordance. *b*) *Functions*: The function of the object in relation to the affordance. *c*) *Appearance*: The visual features of the object that are related to its actions or functions. *d*) *Environment*: The context or environment in which interactions between the object and agent are possible.

We prompt the LLM to select representative phrases for each perspective, denoted as $\phi \in \Phi = \{\text{act}, \text{fun}, \text{app}, \text{env}\}$. Then, we sample phrases $\mathcal{S}_\phi \sim \text{Phrases}(\phi)$ and arrange them to create an enhanced description. The sampling strategy encourages that the generated descriptions maximize the normalized intra-class variance $\tilde{\mathcal{V}}(\mathcal{S})$ and inter-class separability $\tilde{\mathcal{U}}(\mathcal{S})$, thereby aligning with the goals of the scoring function. Design details and experimental results in Section 4.4.3 demonstrate that our approach effectively balances intra-class variance and inter-class separability, enhancing the diversity and distinctiveness of the generated descriptions while preserving their high relevance to the target affordance.

3.3 Feature Extraction Network

3.3.1 Point Cloud Geometric Network. Following the existing 3D affordance detection methods [23, 34, 40], we propose a geometric network based on PointNet++ [36], a popular 3D backbone to extract point-wise embedding with down & up transitions and bridge connection. PointNet++ employs a hierarchical feature leaning architecture in its encoding phase, where each level (i.e. *set abstraction*) has three key operations: *Sampling*, *Grouping*, and *mini-PointNet*. We denote these operations as an encoder layer in Fig.2 (a). One step further, inspired by Point-BERT [50], we introduce a cross-attention mechanism to model the relationship between sets of points, resulting in the refined extraction of geometric features. For the input point cloud $P \in \mathbb{R}^{n\times 3}$, the encoder layer first divides n points into k sets using the furthest distance sampling (*FPS*) and k-nearest-neighbor algorithm (kNN). Thus the position embedding for each set could also be learned from its *FPS* coordinate through a multilayer perceptron (MLP) layer, denoted as $X = \{x_i\}_{i=1}^k$. We then extract the set-wise embeddings via *mini-PointNet*, denoted as $F = \{f_i\}_{i=1}^k$. These positions and embeddings are added element-wisely to form $Z = X \oplus F$, where $Z \in \mathbb{R}^{k\times d}$, and then fed into the L-block Transformer encoder as shown in Fig.2 (d). The process of each block $\phi^{(l)}$ is composed as follows:

$$\tilde{z}_I^{(l),i} = \text{MSA}_l\left(z_I^{(l-1),i}\right) + z_I^{(l-1),i} \tag{6}$$

$$z_I^{(l),i} = \text{MLP}_l\left(\text{LN}\left(\tilde{z}_I^{(l),i}\right)\right) + \tilde{z}_I^{(l),i} \tag{7}$$

where $\text{MSA}_l(\cdot)$ is the multi-head self-attention mechanism at the l-th block, calculated by the softmax-weighted interactions among the input query, key, and value tokens, obtained by three different learnable linear projection weights. Specifically, the self-attention is computed as $\text{Attn}(Q,K,V) = \text{Softmax}(QK^\top/\sqrt{d})V$, where $Q = ZW_Q$, $K = ZW_K$, and $V = ZW_V$. $\text{MLP}_l(\cdot)$ is the multi-layer perceptron at the l-th block, and $\text{LN}(\cdot)$ is Layer Normalization.

We adopt the residual connection to combine the original feature F and its enhanced version to form the final output F'. As shown in Fig.2 (a), we employ multiple encoder layers to continuously expand the level of set abstraction (each followed by a cross-attention module), and propagate these features to the up transition phase, ensuring that the features are effectively distributed across different levels of decoding.

3.3.2 Text Semantic Encoder. To mitigate the confusion of label-based affordance understanding, we adopt a more detailed approach as introduced in Section 3.2. By modeling the diversity and distinctiveness, the FTE module selects key descriptions for each affordance from different perspectives to pre-build a corpus. Then in the training step, given a series of labels $\{l_i\}_{i=1}^n$ which should align the points of the input object $\{p_i\}_{i=1}^n$, we query the corpus database to find relevant phrases for each label and then combine these phrases as a new description $\{t_i\}_{i=1}^n$, just as shown in Fig.2 (b). For instance, given a label *grasp*, we sample perspectives *function* & *action*, and in each we sample only one phrase. They are *Enables manual manipulation* and *Hold object with hands*, respectively. This means that for an object with *grasp* affordance, more precisely a point, the possible function it can have is *Enables manual manipulation* and the potential action that can occur at that point is *Hold object with hands*. Without loss of generality and considering the brevity of the phrase (even for phrase combinations), we extract the whole semantics of the concatenated string through the pretrained text encoder of the CLIP model [37] due to its powerful capability in textual comprehension.

3.4 Contrastive Representation Alignment

Since the geometric network and text encoder have prepared embeddings of point cloud and affordance description respectively, we design a multilevel representation alignment approach to bring the related features closer while distancing the unrelated ones. Specifically, PointNet++ utilizes a stack of decoder layers to integrate learned features from an abstract to concrete manner. As shown in Fig.2 (c), $g_{l,i}$ represents the embedding of the i-th set of points at the l-th level. After the deep layer, point-wise embedding should be extracted, while for the intermediate and shallow layer, the growing region or set-wise embedding could be obtained, denoted as

$\{g_{1,i}\}_{i=1}^{n_1}$, $\{g_{2,i}\}_{i=1}^{n_2}$, and $\{g_{3,i}\}_{i=1}^{n_3}$, where $n_1 > n_2 > n_3$. For each level, we first apply the following *Weighted Cross-Entropy Loss*:

$$\mathcal{L}_{ce}^{l} = \frac{1}{BN}\sum_{i=1}^{BN}\left\{-w_{\mathrm{y_i}}\log\frac{\exp\left(s(g_i, t_{\mathrm{y_i}})/\tau_1\right)}{\sum_{k=1}^{M}\exp(s(g_i, t_k)/\tau_1)}\right\} \tag{8}$$

For brevity, here we use g_i and t_i to denote the geometric embedding of the i-th point (or set) at the l-th level, and the text embedding produced by the proposed FTE, respectively. $s(\cdot,\cdot)$ is the cosine similarity. y_i represents the affordance of GT reference and $w_{\mathrm{y_i}}$ is the weighting coefficient for mitigating affordance imbalance during training. B is the *batch size* and τ is a learnable temperature parameter for controlling the model's certainty and exploration.

To capture the relationships between samples and learn discriminative, generalizable representations, we introduce *Supervised Contrastive Loss* to affordance detection. Specifically, for each available affordance in a *Batch*, the core idea is to first distinguish the positive and negative samples, and then minimize the semantic distance from the positive samples to the reference affordance while pushing negative samples away from the reference affordance. This formulation is as follows:

$$\mathcal{L}_{sc}^{l} = \frac{1}{|\mathcal{M}|}\sum_{t_i\in\mathcal{M}}\left\{-\frac{1}{|\mathcal{X}_i|}\sum_{g_j\in\mathcal{X}_i}\log\frac{\exp\left(s(t_i, g_j)/\tau_2\right)}{\sum_{k=1}^{BN}\exp(s(t_i, g_k)/\tau_2)}\right\} \tag{9}$$

where $\mathcal{X}_i$ denotes the positive samples set for the i-th available affordance and $|\mathcal{X}_i|$ is its cardinality. For a specific *Batch* with $B \cdot N$ elements, it may not have all the affordances so the available affordances set is $\mathcal{M}$. The meanings of other symbols are similar to those in (8).

For the point-wise embedding of the deep layer, the GT reference is obvious while for other layers, it seems to be bewildering. In this case, we record the point sets obtained by PointNet++ during the multisampling phase and analyze the most dominant affordances that should be presented based on point-wise annotations, which allows our design to seamlessly adapt to different levels of abstraction. The final loss function considers the alignment of region-wise and point-wise representations and is formulated as follows:

$$\mathcal{L} = \sum_{l}(\mathcal{L}_{ce}^{l} + \lambda\cdot\mathcal{L}_{sc}^{l}) \tag{10}$$

where λ is a coefficient to balance the two different losses.

4 Experimental Evaluation

This section presents a series of experiments to evaluate the effectiveness of our model.We assess its zero-shot detection capabilities to examine how well it generalizes to unseen affordances. Additionally, we conduct ablation studies to investigate other key aspects of the proposed model. Finally, we validate the significance of accurate affordance understanding for safe operation through real-world robotic experiments.

4.1 Dataset and Baselines

Our experiments are conducted on the 3D AffordanceNet dataset [11], the largest dataset designed for affordance understanding with 3D point cloud data. It consists of 22,949 instances distributed across 23 object categories and 18 affordance labels. We evaluates our methods on the *Label-as-Query* and *Question-as-Query* tasks, with the related test datasets provided by OpenAD [34] and LASO [23]. Label-as-Query is designed to evaluate the model's zero-shot detection ability for multiple extra labels. Question-as-Query is to evaluate the zero-shot performance for 3D objects guided by situational questions. We adapt this dataset by determining whether each point embedding is closer to the question or the background, in order to generate a mask for accurate evaluation.

We compare our method with state-of-the-art approaches for open-vocabulary 3D affordance detection under zero-shot setting: 3DGenZ [29], ZSLPC [6], OpenAD [34], KD-TPC [40] and LASO [23]. Among them, 3DGenZ and ZSLPC are strong baselines for 3D zero-shot learning, while OpenAD and KD-TPC are state-of-the-art methods for open vocabulary affordance detection, incorporating pretrained text encoders and geometric knowledge distillation. LASO is the most recent language-guided affordance segmentation method with focus on complex situational queries. For fair comparison, following OpenAD and LASO evaluation protocols, we use these metrics: For label-as-query task: mIoU (mean Intersection over Union), Acc (Accuracy), and mAcc (mean Accuracy). For question-as-query task: mIoU, AUC (Area Under Curve), SIM (Cosine Similarity), and MAE (Mean Absolute Error).

4.2 Implementation Details

We adopt three encoder-decoder layers in PointNet++, with the alignment module operating at three hierarchical levels using $n_1 = 2048$, $n_2 = 512$, and $n_3 = 128$ point sets, respectively. Both text and geometric embeddings are projected to a shared space of dimension $d = 512$. The loss balance weight is set to $\lambda = 0.25$, and the temperature parameter τ is initialized as $\ln(1/0.07)$. During training, the CLIP text encoder is kept frozen, and only the point cloud branch is optimized using the Adam optimizer with a learning rate of 0.001 and a batch size of 16. All experiments are conducted on an NVIDIA RTX A4500 GPU (20GB).

4.3 Main Results

4.3.1 *Qualitative Results.*

Multi-label as query. Fig. 3 demonstrates the robustness of our model in handling unseen affordances. The proposed method leverages multilevel alignment and cross-attention to enhance discriminative understanding. Compared with OpenAD and KD-TPC, our model produces cleaner predictions with more distinct boundaries, particularly in local regions such as the body of a *vase* and the screen of a *display*. While certain challenges remain in extremely small areas (e.g., the tip of a *knife*), the primary affordance regions are accurately identified.

Question as query. Fig. 4 highlights our model's ability to localize affordance regions in response to diverse functional queries. For example, in (a), when prompted to identify *the point from which water would flow out of the vase* as opposed to *non-functional points*, the model accurately highlights the correct region. This result underscores the importance of functional text descriptions in enabling effective zero-shot inference.

4.3.2 *Quantitative Results.*

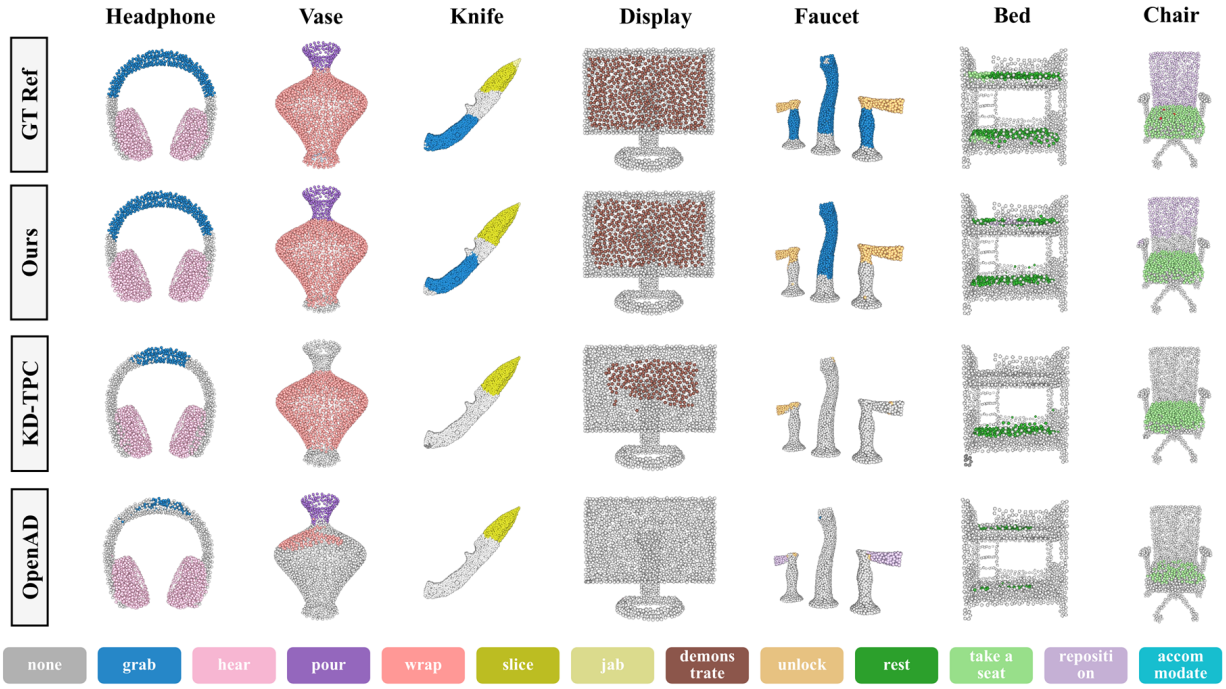


Figure 3: Qualitative comparison of affordance detection results across different methods under the label-as-query setting.

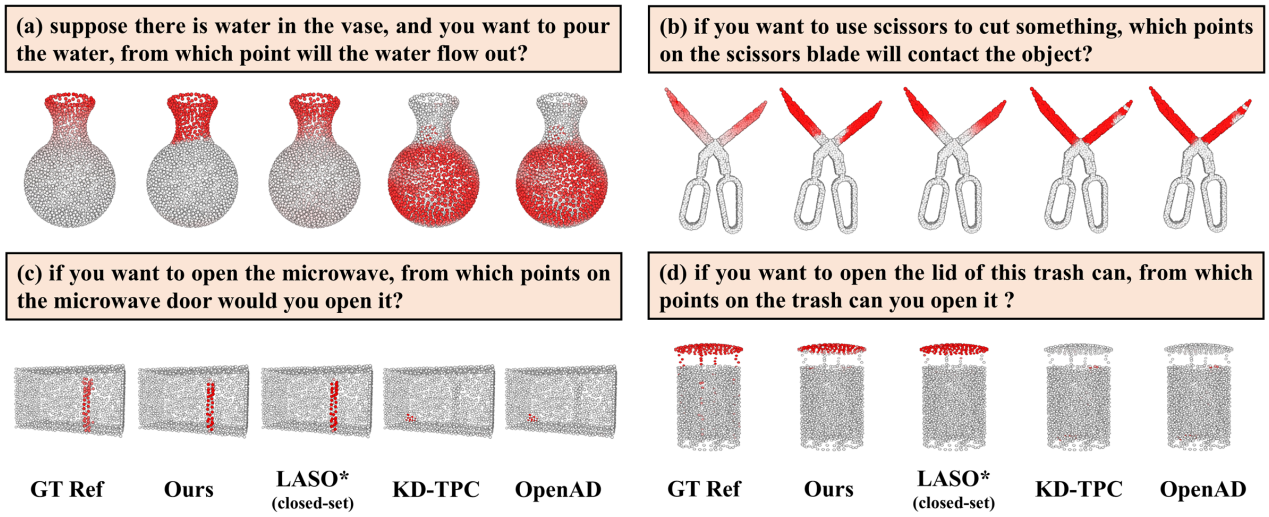


Figure 4: Qualitative comparison of affordance detection results across different methods under the question-as-query setting.

Multi-label as query. Table 1 presents the results on both *partial* and *full-view* point cloud observations, highlighting the robustness and generalization capability of our method across varying levels of input completeness. (a) In the full-view setting, our model achieves the best performance across all evaluation metrics, surpassing the previous state-of-the-art by 14% in detection accuracy and 7% in mIoU. Even without cross-attention, the lightweight variant achieves competitive results, demonstrating the effectiveness and efficiency of the overall design. (b) Partial views are common in real-world scenarios, where robots often perceive objects from limited or occluded viewpoints. In this setting, our model maintains strong performance, improving mIoU by 6% over the best baseline, and demonstrating effective generalization to incomplete observations. Here, results for 3DGenZ [29] and ZSLPC [6] are taken from KD-TPC [40], while all other baselines are re-implemented under the same settings.

Table 1: Label as query: zero-shot detection results

Task	Method	mIoU ↑	Acc ↑	mAcc ↑	Params (M)
Full-view	3DGenZ	0.0646	0.4547	0.1833	1.79
	ZSLPC	0.0997	0.4013	0.1870	1.96
	OpenAD	0.1437	0.4631	0.1951	1.80
	KD-TPC	0.2233	0.4972	0.3429	0.78
	Ours (w/o CA)	0.2653	0.5941	0.4501	0.92
	Ours	**0.2942**	**0.6078**	**0.4829**	3.20
Partial-view	3DGenZ	0.0603	0.4524	0.1586	1.79
	ZSLPC	0.0952	0.4091	0.1716	1.96
	OpenAD	0.1250	0.4525	0.1737	1.80
	KD-TPC	0.2048	0.4872	0.3286	0.78
	Ours	**0.2615**	**0.6020**	**0.4105**	3.20

Table 2: Question as query: zero-shot detection results

Method	mIoU ↑	AUC ↑	SIM ↑	MAE ↓	Params (M)
LASO* (closed-set)	0.1995	0.8527	0.6080	0.1023	9.10
OpenAD	0.1026	0.5968	0.2251	**0.1827**	1.80
KD-TPC	0.1083	0.6066	0.3372	0.2563	0.78
Ours (w/o CA)	0.1218	0.6153	0.3467	0.2760	0.92
Ours	**0.1315**	**0.6216**	**0.3558**	0.2716	3.20
Ours* (linear probe)	0.1756	0.8438	0.6129	0.1078	3.20

Question as query. Table 2 demonstrates that our method achieves the best performance on most metrics under the zero-shot setting. Specifically, it obtains the highest mIoU, surpassing OpenAD by 2.91% and KD-TPC by 2.35%. It also outperforms all baselines in AUC and SIM, indicating superior localization quality and region-level consistency. For reference, we include the performance of LASO* (closed-set), a language-aware affordance detection model trained with full supervision and evaluated in a closed-set setting. As such, its performance serves as an upper bound.

To assess the generalizability of our point cloud encoder, we freeze its parameters and attach a classification head trained in the closed set, denoted as Ours* (linear probe). Compared to LASO*, our model achieves comparable AUC and SIM scores despite using only one-third of the parameters, with a moderate drop in mIoU. This highlights the strong representational capacity and efficiency of our learned features.

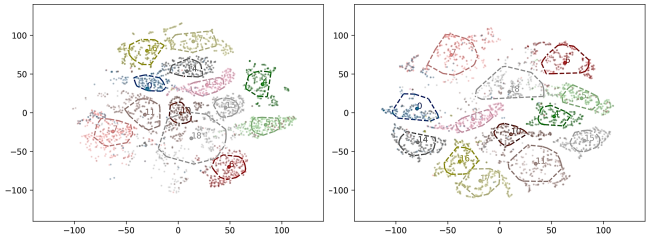
4.4 Ablation Studies

4.4.1 Effect of Proposed Components. Table 3 shows the progressive impact of each proposed component. The results comprehensively validate the effectiveness of function text enhancement (FTE) and supervised contrastive loss (SC), with FTE particularly excelling in generalization and SC in discrimination. It also suggests the combined use of multilevel alignment (ML) and cross attention (CA), which may derive from the fact that ML captures multiple levels of locality, while CA relates locality to the global context.

4.4.2 Visualization of Learned Geometric Embeddings. To compare the effects of label-based and FTE-based descriptions on the learned embeddings, we conduct t-SNE visualizations for OpenAD [34] and our method. Specifically, we collect $N = 500$ samples per affordance class and project their 512-D embeddings into 2-D using t-SNE. For

Table 3: Effect of the proposed components

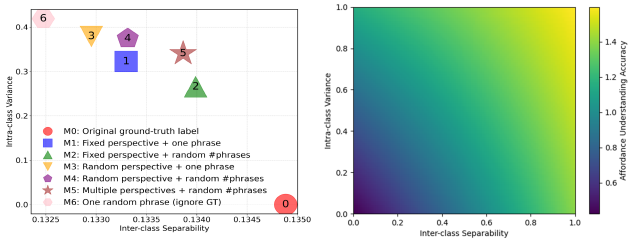
Variants	Label as Query			Question as Query			
	mIoU ↑	Acc ↑	mAcc ↑	mIoU ↑	AUC ↑	SIM ↑	MAE ↓
Baseline	0.1531	0.4619	0.2354	0.1033	0.5915	0.2124	0.1802
+SC	0.1963	0.4963	0.3402	0.1088	0.6084	0.3709	0.3374
+FTE	0.2504	0.5760	0.3829	0.1168	0.6150	0.3638	0.2961
+SC+FTE	0.2523	0.5818	0.4377	0.1183	0.6131	0.3671	0.3073
+SC+FTE+ML	0.2653	0.5941	0.4501	0.1218	0.6153	0.3467	0.2760
+SC+FTE+ML+CA	0.2942	0.6078	0.4829	0.1315	0.6216	0.3558	0.2716



(a) t-SNE visualization of geometric embeddings from *OpenAD*. (compact, overlapping regions)

(b) t-SNE visualization of geometric embeddings from the proposed *Aff3DFunc*. (distinct boundaries, larger spread)

Figure 5: t-SNE visualizations for geometric embeddings.



(a) Trade-off between metrics across different functional text construction strategies.

(b) Heatmap showing how intra-class variance and inter-class separability jointly affect accuracy.

Figure 6: Further analysis of functional text strategies.

clearer comparison, we compute the class centers and use the kNN algorithm to select the $N/2$ points nearest to each center, from which we construct convex hulls. As shown in Fig.5, our method yields more distinct category boundaries and greater intra-class spread, reflecting the effects of FTE in enhancing both separability and diversity.

4.4.3 Further Discussion on Functional Text. As introduced in Section 3.2, our FTE module is based on the Information Bottleneck (IB) principle, aiming to generate affordance descriptions that balance semantic diversity and inter-class separability by sampling and composing concepts from a concept pool. We conduct detailed analyses of key design choices, as summarized in Table 4. **(a) Choice of LLM.** We evaluate different LLMs for generating the concept pool. Results show only minor performance variations, confirming the robustness of FTE. ChatGPT-3.5 is adopted in the main experiments due to its balance of availability and performance. **(b) Prompting Strategy.** FTE outperforms alternatives relying on raw labels or simple prompts, producing richer and more discriminative semantics. **(c) Sampling and Composition.** We compare six variants based on different semantic perspectives and sampling

granularities (see Fig.6 (a) for details). The optimal configuration (*M5*) achieves a favorable balance between intra-class variance and inter-class separability, effectively expanding the semantic space while preserving clear category boundaries. To explore the relationship among $\mathcal{V}$, $\mathcal{U}$, and detection accuracy $\mathcal{Y}$, we normalize both metrics and model their interaction using least squares fitting. As shown in Fig.6 (b), the observed negative interaction indicates that simultaneous increases in both metrics attenuate the growth rate of accuracy, highlighting the tension between variance and separability. **(d) Phrase Encoding.** We compare two encoding methods: concatenating sampled phrases before feeding them into the text encoder versus encoding individually followed by pooling. The concatenation strategy not only yields better downstream performance but also aligns with our theoretical expectations, as this fusion variant significantly reduces both normalized variance (from 0.58 to 0.29) and inter-class separability (from 0.34 to 0.13). **(e) Pool Size.** The pool size denotes the number of LLM-generated descriptions per affordance. As evidenced in Table 5, small pool sizes lead to constrained intra-class variance and limited semantic coverage. Increasing pool size improves diversity but slightly reduces separability. Beyond a threshold, marginal gains diminish, indicating minimal benefit from excessively large pools.

Table 4: Comparison of different strategies for FTE

Category	Method	Label as Query			Question as Query		
		mIoU ↑	Acc ↑	mAcc ↑	mIoU ↑	AUC ↑	MAE ↓
LLM	GPT 3.5	0.2942	0.6078	0.4829	0.1315	0.6216	0.2716
	GPT 4o	0.2895	0.6103	0.4648	0.1313	0.6360	0.3508
	DeepSeek R1	0.2795	0.6260	0.5085	0.1264	0.6255	0.2983
Prompting	Label Only	0.1741	0.4874	0.2997	0.1137	0.6054	0.1917
	Simple Prompt	0.2145	0.4852	0.3422	0.1207	0.6187	0.2559
	Proposed FTE	0.2942	0.6078	0.4829	0.1315	0.6216	0.2716
Sampling	M0	0.1812	0.4871	0.3202	0.1229	0.6104	0.1999
	M1	0.1953	0.3650	0.3946	0.1055	0.6056	0.4486
	M2	0.2302	0.4113	0.4926	0.0889	0.5894	0.3037
	M3	0.2482	0.5396	0.4305	0.1265	0.6362	0.3398
	M4	0.2797	0.5899	0.4879	0.1179	0.6210	0.3274
	M5	0.2942	0.6078	0.4829	0.1315	0.6216	0.2716
Phrase Fusion	Fusion	0.2747	0.5637	0.4452	0.1226	0.6375	0.3808
	Concat	0.2942	0.6078	0.4829	0.1315	0.6216	0.2716

Table 5: Pool size

Pool Size	Variance	Separability
4	0.2889	0.1339
64	0.3463	0.1331
100	0.3537	0.1329

Table 6: Inference efficiency

Config	FLOPs (G)	Params (M)	Latency (ms)
Ours	5.37	3.20	104.4
Ours w/o CA	4.83	0.92	102.8
Ours w/o ML	5.33	3.07	104.1

4.5 Robotics Application

To validate the proposed affordance detection method in robotic applications, we conducted real-world experiments using a Unitree GO2 mobile platform equipped with a 6-DoF D1 robotic arm and a parallel gripper. We focus on safety-critical manipulation tasks [4] that require precise affordance understanding, such as distinguishing between a knife's handle (graspable) and blade (hazardous). A diverse set of household objects was used for evaluation.

As illustrated in Fig.7 (1) and (3), our method successfully identified functional regions, enabling reliable pick-and-place operations.

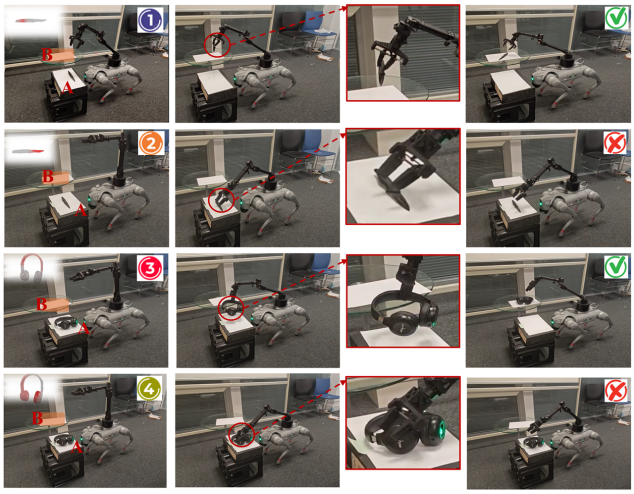


Figure 7: Robotics manipulation validation.

In contrast, OpenAD's predictions (Fig.7 (2) and (4)) often targeted non-functional or unsafe regions (e.g., knife blades, headphone earcups), triggering emergency stops during execution. These experiments demonstrate the effectiveness of our method in real-world robotic scenarios and underscore the critical role of accurate affordance detection in ensuring operational safety.

Efficiency is also a key factor in real-world robotic applications. In this context, methods such as 3DAffordanceLLM [8] and SeqAfford [49] rely on large language models or multimodal backbones for semantic understanding, resulting in substantial computational overhead and limited deployability. Table 6 reports the efficiency of our proposed approach. All evaluations were conducted on an NVIDIA RTX A4500 GPU (20GB), demonstrating that our method supports real-time inference with minimal computational cost and is readily deployable in latency-sensitive settings.

5 Conclusion and Limitation

In this paper, we present *3DAffFunc*, a lightweight open-vocabulary framework for 3D affordance understanding. The proposed method incorporates functional text enhancement to enrich affordance semantics and employs multimodal encoders to jointly extract geometric and semantic features. To enhance cross-modal interaction, we introduce a multilevel representation alignment strategy that establishes robust correlations across multiple abstraction levels. Extensive experiments demonstrate that *3DAffFunc* significantly advances 3D affordance understanding in zero-shot scenarios.

Although *3DAffFunc* shows notable improvements, 3D affordance understanding remains a challenging problem, with current state-of-the-art mIoU still below 0.3. Our model can reliably identify primary affordances but struggles with fine-grained regions, such as the tip of a knife in the "jab" affordance (Fig. 3). In addition, the CLIP-based text encoder is constrained by limited context length, and our robotic experiments are currently qualitative. Future work will explore incorporating priors from foundation models [1, 28, 51], investigating more flexible language encoders [2, 24, 32], and developing comprehensive benchmarks for real-world evaluation.

Acknowledgments

This work was supported by the Graduate School Scholarship from the College of Science and Engineering, University of Glasgow.

References

[1] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. 2025. Generalized Few-shot 3D Point Cloud Segmentation with Vision-Language Model. *arXiv preprint arXiv:2503.16282* (2025).

[2] Mothilal Asokan, Kebin Wu, and Fatima Albreiki. 2025. FineLIP: Extending CLIP's Reach via Fine-Grained Alignment with Longer Text Inputs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14495–14504.

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

[4] Lukas Brunke, Yanni Zhang, Ralf Römer, Jack Naimer, Nikola Staykov, Siqi Zhou, and Angela P Schoellig. 2025. Semantically safe robot manipulation: From semantic scene understanding to motion safeguards. *IEEE Robotics and Automation Letters* (2025).

[5] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. 2023. Mixed autoencoder for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22742–22751.

[6] Ali Cheraghian et al. 2019. Zero-shot learning of 3D point cloud objects. In *Int. Conf. on Machine Vision Applications (MVA)*. 1–6.

[7] Hengshuo Chu, Xiang Deng, Xiaoyang Chen, Yinchuan Li, Jianye Hao, and Liqiang Nie. 2025. 3D-AffordanceLLM: Harnessing Large Language Models for Open-Vocabulary Affordance Detection in 3D Worlds. *arXiv preprint arXiv:2502.20041* (2025).

[8] Hengshuo Chu, Xiang Deng, Qi Lv, Xiaoyang Chen, Yinchuan Li, Jianye HAO, and Liqiang Nie. [n. d.]. 3D-AffordanceLLM: Harnessing Large Language Models for Open-Vocabulary Affordance Detection in 3D Worlds. In *The Thirteenth International Conference on Learning Representations*.

[9] Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2024. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 1274–1285.

[10] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. 2024. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14531–14542.

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

[12] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. 2020. Action Modifiers: Learning From Adverbs in Instructional Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Hazel Doughty, Fida Mohammad Thoker, and Cees GM Snoek. 2024. Locomotion: Learning motion-focused video-language representations. In *Proceedings of the Asian Conference on Computer Vision*. 50–70.

[14] Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, and Bin Zhao. 2024. Learning 2d invariant affordance knowledge for 3d affordance grounding. *arXiv preprint arXiv:2408.13024* (2024).

[15] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. 2023. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7081–7091.

[16] James Jerome Gibson. 1966. The senses considered as perceptual systems. (1966).

[17] Mohammed et al. Hassanin. 2021. Visual affordance and function understanding: A survey. *Comput. Surveys* 54, 3 (2021), 1–35.

[18] Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. 2024. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

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

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

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

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[23] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. 2024. LASO: Language-guided Affordance Segmentation on 3D Object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14251–14260.

[24] Xiaoran Liu, Ruixiao Li, Mianqiu Huang, Zhigeng Liu, Yuerong Song, Qipeng Guo, Siyang He, Qiqi Wang, Linlin Li, Qun Liu, et al. 2025. Thus spake long-context large language model. *arXiv preprint arXiv:2502.17129* (2025).

[25] Zichuan Liu, Zefan Wang, Linjie Xu, Jinyu Wang, Lei Song, Tianchun Wang, Chunlin Chen, Wei Cheng, and Jiang Bian. 2024. Protecting your llms with information bottleneck. *Advances in Neural Information Processing Systems* 37 (2024), 29723–29753.

[26] Liangsheng Lu, Wei Zhai, Hongchen Luo, Yu Kang, and Yang Cao. 2022. Phrase-based affordance detection via cyclic bilateral interaction. *IEEE Transactions on Artificial Intelligence* 4, 5 (2022), 1186–1198.

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

[28] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. 2024. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems* 37 (2024), 76819–76847.

[29] Björn et al. Michele. 2021. Generative zero-shot learning for semantic segmentation of 3D point clouds. In *Int. Conf. on 3D Vision (3DV)*. 992–1002.

[30] Lorenzo Mur-Labadia, Jose J. Guerrero, and Ruben Martinez-Cantin. 2023. Multi-label Affordance Mapping from Egocentric Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5238–5249.

[31] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. 2015. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1374–1381.

[32] Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M Asano, Nanne Van Noord, Marcel Worring, and Cees GM Snoek. [n. d.]. TULIP: Token-length Upgraded CLIP. In *The Thirteenth International Conference on Learning Representations*.

[33] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. 2024. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704* (2024).

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

[35] Pengxiang Ouyang, Jianan Chen, Qing Ma, Zheng Wang, and Cong Bai. 2024. Distinguishing Visually Similar Images: Triplet Contrastive Learning Framework for Image-text Retrieval. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30 (2017).

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

[38] Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. 2025. Enhancing Cognition and Explainability of Multimodal Foundation Models with Self-Synthesized Data. *arXiv preprint arXiv:2502.14044* (2025).

[39] Ramesh Ashok Tabib, Dikshit Hegde, and Uma Mudenagudi. 2024. LGAfford-Net: A Local Geometry Aware Affordance Detection Network for 3D Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5261–5270.

[40] Tuan Van Vo, Minh Nhat Vu, Baoru Huang, Toan Nguyen, Ngan Le, Thieu Vo, and Anh Nguyen. 2024. Open-vocabulary affordance detection using knowledge distillation and text-point correlation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13968–13975.

[41] Xin Wang, Hong Chen, Zihao Wu, Wenwu Zhu, et al. 2024. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[42] Shuhuan Wen, Tao Wang, and Sheng Tao. 2022. Hybrid CNN-LSTM architecture for LiDAR point clouds semantic segmentation. *IEEE Robotics and Automation Letters* 7, 3 (2022), 5811–5818.

[43] Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. 2024. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *Advances in Neural Information Processing Systems* 36 (2024).

[44] Lingjing Xu, Yang Gao, Wenfeng Song, and Aimin Hao. 2024. Weakly Supervised Multimodal Affordance Grounding for Egocentric Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6324–6332.

[45] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. 2023. Grounding 3D Object Affordance from 2D Interactions in Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10905–10915.

[46] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. 2024. LEMON: Learning 3D Human-Object Interaction Relation from 2D Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16284–16295.

[47] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. 2024. EgoChoir: Capturing 3D Human-Object Interaction Regions from Egocentric Views. *arXiv preprint arXiv:2405.13659* (2024).

[48] Luo Yiyang, Ke Lin, and Chao Gu. 2024. Context-Aware Indoor Point Cloud Object Generation through User Instructions. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10182–10190.

[49] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. 2025. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1691–1701.

[50] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 2022. Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19291–19300.

[51] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2023. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems* 36 (2023), 45533–45547.

[52] Huiming Zheng, Wei Gao, Zhuozhen Yu, Tiesong Zhao, and Ge Li. 2024. Viewpcgc: view-guided learned point cloud geometry compression. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7152–7161.