

Task-Aware 3D Affordance Segmentation via 2D Guidance and Geometric Refinement

Lian He^{1,2}, Meng Liu^{2,3*}, Qilang Ye^{4,2}, Yu Zhou^{4,2}, Xiang Deng⁵, Gangyi Ding^{1*}

¹ School of Computer Science & Technology, Beijing Institute of Technology

² Zhongguancun Academy

³ Shandong Jianzhu University

⁴ VCIP & TMCC & DISSec, College of Computer Science & College of Cryptology and Cyber Science, Nankai University

⁵ Harbin Institute of Technology, Shenzhen

Abstract

Understanding 3D scene-level affordances from natural language instructions is essential for enabling embodied agents to interact meaningfully in complex environments. However, this task remains challenging due to the need for semantic reasoning and spatial grounding. Existing methods mainly focus on object-level affordances or merely lift 2D predictions to 3D, neglecting rich geometric structure information in point clouds and incurring high computational costs. To address these limitations, we introduce **Task-Aware 3D Scene-level Affordance segmentation (TASA)**, a novel geometry-optimized framework that jointly leverages 2D semantic cues and 3D geometric reasoning in a coarse-to-fine manner. To improve the affordance detection efficiency, TASA features a task-aware 2D affordance detection module to identify manipulable points from language and visual inputs, guiding the selection of task-relevant views. To fully exploit 3D geometric information, a 3D affordance refinement module is proposed to integrate 2D semantic priors with local 3D geometry, resulting in accurate and spatially coherent 3D affordance masks. Experiments on SceneFun3D demonstrate that TASA significantly outperforms the baselines in both accuracy and efficiency in scene-level affordance segmentation.

Introduction

Affordances describe the latent functional properties of objects, such as handles, knobs, and switches, which allow specific actions (Gibson 2014). Understanding affordances in 3D environments is essential for enabling embodied agents to interact effectively with the physical world (Radford et al. 2021). To perform goal-directed behaviors, agents must not only perceive these functional components but also ground them with high-level instructions, often expressed through natural language. This grounding is especially challenging in cluttered environments, where language instructions implicitly reference specific, functionally relevant object parts. For example, executing the command “open the bottom drawer” requires identifying and localizing the correct handle among many visually similar alternatives. Solving this task demands: 1) *commonsense reasoning to infer implied*

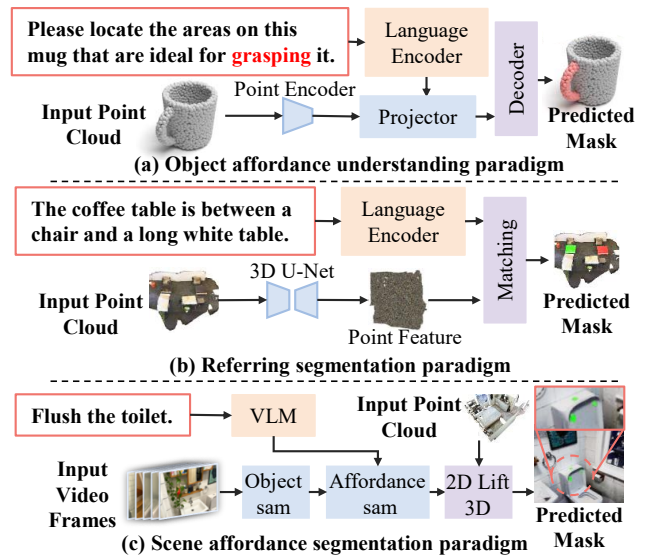


Figure 1: Comparison among (a) object affordance understanding, (b) referring segmentation, and (c) scene affordance segmentation paradigms.

functional targets from language, and 2) *fine-grained spatial perception to detect small and context-sensitive elements*. Effectively bridging natural language and 3D perception under such conditions is a critical step toward building intelligent agents capable of operating autonomously in real-world environments.

Despite recent progress, existing approaches to 3D affordance understanding remain limited in several critical aspects. First, most methods adopt a narrow focus by targeting single-object affordance detection in isolation (Chu et al. 2025; Yang et al. 2023; Nguyen et al. 2023; Van Vo et al. 2024). For example, they identify grasp points on a mug (Figure 1(a)) without considering the broader scene context. These object-centric paradigms, which treat individual objects as standalone inputs, **fail to scale to complex, scene-level reasoning** where affordances often depend on inter-object relationships and spatial arrangements. Second, many approaches rely heavily on 2D image-based informa-

*Corresponding Authors

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tion, typically projecting 2D segmentation masks into 3D via camera poses. This lifting process often overlooks essential geometric cues. As shown in Figure 1(c), when handling the “flush the toilet” task, it incorrectly segments point clouds with similar textures near the flush button into the affordance region, resulting in blurred segmentation outcomes that fail to differentiate between affordance-critical regions and surrounding areas with similar textures. **The absence of true 3D spatial constraints hampers the accurate localization of functional parts.** Third, current frameworks are inefficient in handling multi-view data. Typical methods for 3D affordance segmentation (Corsetti et al. 2025) require processing hundreds of RGB frames, yet treat all views equally, **resulting in substantial redundancy and computational overhead without exploiting view-specific utility.**

To address the aforementioned challenges, we propose a **Task-Aware 3D Scene-level Affordance Segmentation Framework (TASA)**, a novel geometry-optimized framework that jointly leverages 2D semantic cues and 3D geometric reasoning in a coarse-to-fine manner. To our knowledge, this is the first geometry-optimized framework explicitly designed for affordance understanding in 3D scenes.

First, to move beyond the single-object paradigm, our method operates directly on full-scene 3D environments. It leverages vision-language models (VLMs) to parse task instructions and extract scene-level semantic cues from both 3D point clouds and multi-view RGB images. This enables context-aware affordance reasoning across objects, accounting for inter-object relationships and spatial configurations.

Second, to reduce the inefficiency of uniform multi-view processing, we introduce a CLIP-based affordance-weighted frame selector. Guided by VLM-extracted task elements, it prioritizes views that are semantically aligned with the action, e.g., focusing on cabinet handles instead of irrelevant views like a “TV”, for the instruction “open the drawer of the cabinet with the TV on top”. This selective strategy significantly reduces redundancy while preserving task-relevant information.

Third, to mitigate over-reliance on 2D projections, we refine 2D segmentation masks by fusing them with 3D geometric features. Given the complexity of directly interpreting high-resolution 3D point clouds, we use 2D masks as a guidance to partition the 3D point cloud, enabling the model to focus on critical regions and achieve precise localization of functional components in texture-similar areas. In addition, we design a multi-objective loss function tailored for high-fidelity point clouds, where traditional methods often fail. This loss encourages fine-grained and spatially consistent segmentation across densely sampled scenes.

We evaluate our method on SceneFun3D, the only existing dataset with 3D affordance segmentation annotations in real-world indoor scenes. The dataset presents multiple challenges: 1) objects with similar visual appearances (e.g., dresser handles), 2) ambiguous and spatially complex task descriptions (e.g., “open the second drawer”), and 3) dense, high-resolution point clouds that demand precise spatial reasoning. Our method demonstrates robust performance across all these scenarios.

Our contributions are summarized as follows:

- We propose a Task-Aware 3D Scene-level Affordance Segmentation Framework that jointly leverages 2D semantic cues and 3D geometric reasoning in a coarse-to-fine manner.
- We design a CLIP-based frame selection strategy guided by VLM-extracted task elements, and a 3D affordance refinement module to produce more precise and spatially consistent affordance masks.
- We achieve state-of-the-art results in 3D scene-level affordance segmentation with high-fidelity point clouds, significantly improving mAP from 7.6 to 23.2 and mIoU from 15.2 to 19.7.

Related Work

Affordance Understanding. Affordance research spans both 2D and 3D domains. In 2D, recent efforts explore off-the-shelf segmentation models and weakly supervised learning (Xu et al. 2024; Li et al. 2024; Xu and Mu 2025; Qian et al. 2024) for efficient RGB-based affordance parsing. However, these approaches lack 3D grounding, limiting their applicability to real-world embodied interaction.

In the 3D domain, most existing methods focus on object-centric affordance prediction (Qian et al. 2024; Chu et al. 2025; Yang et al. 2023; Lu et al. 2025; Shao et al. 2025; Yang et al. 2024). While some recent work attempts open-vocabulary affordance reasoning (Nguyen et al. 2023; Van Vo et al. 2024), these methods rely on isolated-object datasets and struggle to scale to complex scenes with clutter and occlusion. Their limited spatial context restricts fine-grained localization of functional parts in realistic settings. To our knowledge, Fun3DU (Corsetti et al. 2025) is the only method targeting scene-level 3D affordance segmentation. However, it suffers from severe computational inefficiency and merely lifts 2D segmentations into 3D. This leads to cascading errors when initial 2D predictions are inaccurate.

In contrast, our method introduces a task-aware frame selection module that leverages CLIP embeddings guided by VLM-derived affordance cues. This drastically improves efficiency, reducing computation time by orders of magnitude compared to Fun3DU. Crucially, we further enhance segmentation robustness by integrating 3D geometric features into the refinement process, enabling reliable affordance localization in cluttered, high-fidelity environments.

3D Referring Segmentation. Among 3D scene understanding tasks, 3D referring segmentation is partially related to our problem setting. Given a natural language expression, these methods aim to segment the corresponding object in a 3D scene. Existing approaches fall into two main categories: referring expression-based methods (Chen, Chang, and Nießner 2020; Achlioptas et al. 2020; Luo et al. 2022), and superpoint-based segmentation (Wu et al. 2024; Sun et al. 2023; Engelmann et al. 2020). While effective for whole-object localization, these methods lack the granularity needed for actionable manipulation. For example, the instruction “close the bedroom door” requires localizing the doorknob, not the entire door. Furthermore, most models are trained on coarse, low-resolution point clouds, limiting

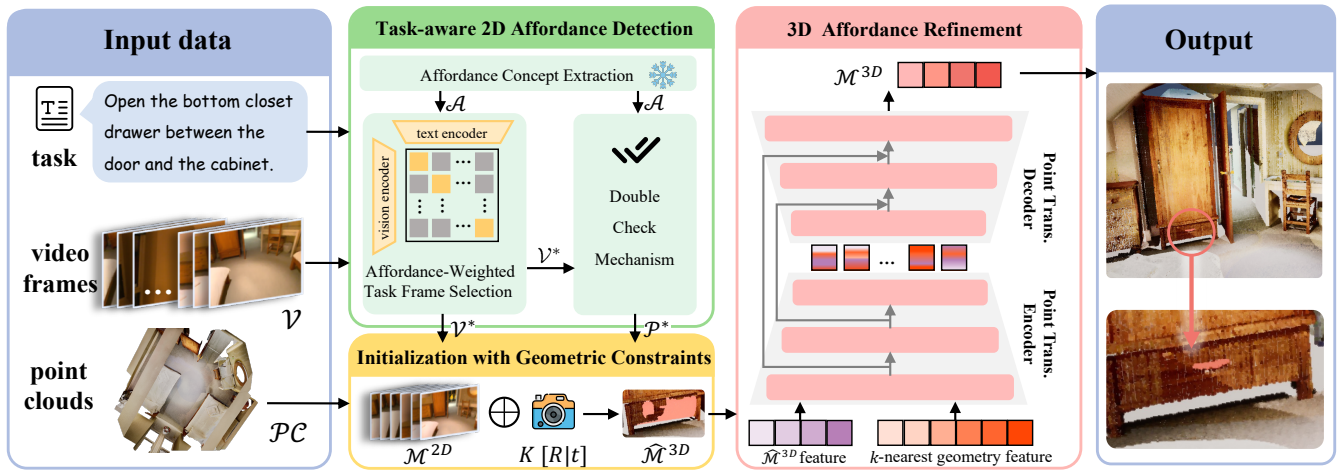


Figure 2: Overview of our Task-Aware 3D Scene-level Affordance Segmentation framework (TASA).

their generalization to high-fidelity 3D data common in real-world robotics applications.

Vision-Language Models. VLMs enable cross-modal reasoning through joint encoding of visual and textual inputs, and can be broadly categorized into 2D and 3D variants. 2D VLMs (Chen et al. 2024; Radford et al. 2021; Li et al. 2022; Alayrac et al. 2022; Liu et al. 2018a,b), have seen rapid progress, driven by large-scale image-text pre-training. These models typically combine a vision encoder with a language model via various fusion architectures. For example, LLaVA (Liu et al. 2023) fuses CLIP (Radford et al. 2021) features into a Vicuna-based language model using a lightweight MLP adapter. In contrast, 3D VLMs are still in their early stages, mainly focusing on grounding generic objects in 3D scenes. Due to the lack of large-scale 3D vision-language datasets and the absence of strong 3D foundation models, current 3D VLMs exhibit limited generalization, particularly in fine-grained affordance reasoning within complex environments.

Method

Overview of the Framework

Figure 2 illustrates our proposed framework for Task-Aware 3D Scene-level Affordance Segmentation, which adopts a coarse-to-fine pipeline to progressively improve segmentation accuracy. The objective is to localize actionable regions in a 3D point cloud based on natural language instructions.

Given a task description \mathcal{T} , a high-precision point cloud $\mathcal{PC} = \{p_i \in \mathbb{R}^3 \mid i = 1, \dots, M\}$, and set of multi-view RGB images $\mathcal{V} = \{v_1, v_2, \dots, v_N\} \in \mathbb{R}^{N \times H \times W \times 3}$, where p_i is individual 3D points, M denotes the total number of points in the point cloud, v_i refers to every RGB image, N is the number of views, H and W are the image height and width, the framework outputs a 3D affordance mask indicating task-relevant regions suitable for downstream robotic manipulation.

Our method comprises two main stages: 1) Task-aware affordance detection, which employs vision-language reason-

ing to guide informative frame selection, and 2) 3D affordance mask refinement, which incorporates geometric features to ensure precise and spatially grounded segmentation.

Task-Aware 2D Affordance Detection

This module extracts reliable 2D manipulable points from natural language instructions and image frames to provide task-aligned guidance for downstream 3D segmentation. It consists of three components: 1) Affordance Concept Extraction using a VLM to identify task-relevant interactive components from text, 2) Affordance-Weighted Frame Selection via a CLIP-based similarity metric modulated by extracted concepts to prioritize informative views, and 3) Manipulable Point Validation through a double-check mechanism that enhances robustness via reverse checks and fallback queries.

Affordance Concept Extraction. Given a natural language task T , we extract a set of interactive concepts essential for completing the instruction. These typically correspond to manipulable object parts (e.g., “handle” in “open the bottom closet drawer between the door and the cabinet”). Formally, let $\mathcal{A} = \{a_1, a_2, \dots, a_k\}$ denote the set of extracted affordance concepts. These are obtained via the VLM:

$$\mathcal{A} = \text{VLM}_{\text{affordance}}(T; \theta_{\text{VLM}}), \quad (1)$$

where θ_{VLM} represents the parameters of the VLM. The output \mathcal{A} encodes functional cues semantically grounded in the instruction.

Affordance-Weighted Frame Selection. Multi-view image streams often contain frames irrelevant to the task (e.g., occluded views, distant background clutter), which can degrade downstream performance and increase computational overhead. To address this, we propose a semantic frame selection strategy that emphasizes frames aligned with extracted affordance concepts.

Standard CLIP-based similarity (Radford et al. 2021) tends to bias toward dominant nouns (e.g., “cabinet”), neglecting functional components like “handle”. To mitigate

this, we introduce an affordance-weighted similarity metric. Let $\text{CLIP}_{\text{img}} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^d$ and $\text{CLIP}_{\text{txt}} : \mathcal{L} \rightarrow \mathbb{R}^d$ be the CLIP encoders for images and text, respectively, with embedding dimension $d = 512$. For each image frame $v_i \in \mathcal{V}$ and task T , we compute $\mathbf{e}_i^{\text{img}} = \text{CLIP}_{\text{img}}(v_i)$, serving as the reference embedding \mathbf{e}^{ref} of frame v_i , $\mathbf{e}_T^{\text{txt}} = \text{CLIP}_{\text{txt}}(T)$, and the baseline CLIP similarity:

$$\text{Sim}_{\text{CLIP}}(v_i, T) = \frac{\mathbf{e}_i^{\text{img}} \cdot \mathbf{e}_T^{\text{txt}}}{\|\mathbf{e}_i^{\text{img}}\|_2 \cdot \|\mathbf{e}_T^{\text{txt}}\|_2} \in [-1, 1], \quad (2)$$

serving as a baseline metric quantifying the overall semantic alignment between frame v_i and task T , capturing contextual information such as dominant entities (e.g., ‘‘cabinet’’) that provide scene-level relevance. To emphasize actionable components, we compute CLIP embeddings for each affordance concept $a_i \in \mathcal{A}$:

$$\mathbf{e}_{a_i}^{\text{txt}} = \text{CLIP}_{\text{txt}}(a_i) \in \mathbb{R}^d. \quad (3)$$

which are directly utilized in the affordance-weighted calculation. The affordance-weighted similarity score S , computed for each frame v_i :

$$\begin{cases} S = \frac{1}{Z} \sum_{a_i \in \mathcal{A}} \exp(\mathbf{e}_{a_i}^{\text{txt}} \cdot \mathbf{e}^{\text{ref}}), \\ Z = \sum_{a_i \in \mathcal{A}} \exp(\mathbf{e}_{a_i}^{\text{txt}} \cdot \mathbf{e}^{\text{ref}}), \end{cases} \quad (4)$$

$\text{Sim}_{\text{CLIP}}(v_i, T)$ and S are complementary and jointly contribute to frame selection. While S prioritizes functional affordance concepts (e.g., handle) critical for task execution. $\text{Sim}_{\text{CLIP}}(v_i, T)$ ensures alignment with the broader task context (e.g., ‘‘cabinet’’), preventing selection of irrelevant frames containing isolated affordance components.

For final frame selection, we calculate $\text{FinalScore}(v_i)$ as:

$$\text{FinalScore}(v_i) = \alpha_a \cdot \text{Sim}_{\text{CLIP}}(v_i, T) + (1 - \alpha_a) \cdot S \quad (5)$$

where $\alpha \in [0, 1]$ balances contextual and affordance priorities. We rank frames by $\text{FinalScore}(v_i)$ and select the top- K frames as \mathcal{V}^* , prioritizing semantically aligned, interactive elements over visually dominant. This method reduces the number of frames processed by focusing on the most relevant ones, enhancing both efficiency and accuracy.

Manipulable Point Validation. For each select frame $v \in \mathcal{V}^*$, we extract 2D manipulable points using a VLM, enhanced by a double-check mechanism to improve both robustness and accuracy.

We first issue a task-specific prompt to the VLM, asking it to identify spatial locations within the image that are critical for completing the given instruction T . This yields a set of candidate points: $\mathcal{P} = \{(x_1, y_1), \dots, (x_m, y_m)\}$, which represent likely affordance regions (e.g., handles or switches) in the task-relevant frames.

However, direct one-shot VLM extraction often suffers from missed detections or false positives due to semantic ambiguity (e.g., the term ‘‘handle’’ may refer to visually diverse components). To address this, we introduce a double-check mechanism composed of two sequential steps: reverse validation and fallback querying, as illustrated in Figure 3.

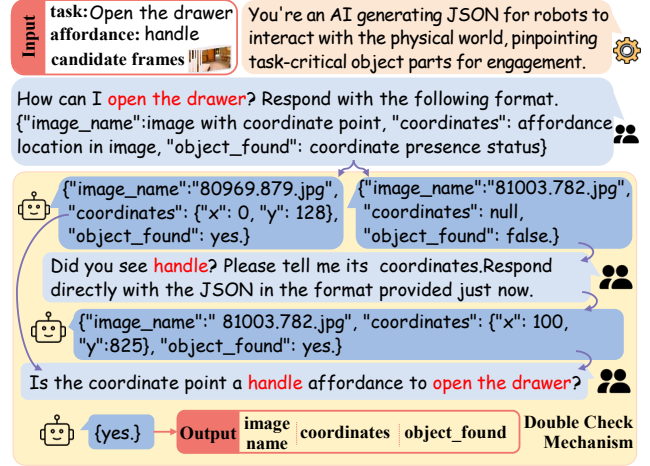


Figure 3: Illustration of the Double-Check Mechanism.

- **Reverse Validation:** For each candidate point $p_i^{2D} = (x_i, y_i) \in \mathcal{P}$, we verify whether interacting with this point is sufficient to complete task T . We define a verification function: $VLM_{\text{verify}}(v_i, p_i^{2D}, T)$, which checks whether point p_i corresponds to a functionally correct component (e.g., ensuring the point lies on a drawer handle rather than on the drawer surface). Points that fail validation are discarded, resulting a validation set of points, denoted as $\mathcal{P}^{\text{valid}}$
- **Fallback Query:** If no candidate point is extracted ($\mathcal{P} = \emptyset$), we trigger a fallback mechanism. For each affordance concept $a_i \in \mathcal{A}$ identified earlier, we query the VLM to estimate relevant 2D locations: $VLM_{\text{point}}(v_i, a_i; \theta_{VLM})$. This yields a supplementary set of points, denoted as \mathcal{P}^* .

The final set of verified manipulable points is the union of the reverse and fallback results (i.e., $\mathcal{P}^{\text{valid}} \cup \mathcal{P}^*$), ensuring both precision and recall in affordance localization. These points serve as coarse spatial cues for subsequent 2D mask generation, forming the initial step in our coarse-to-fine affordance segmentation pipeline.

3D Affordance Refinement

This module fuses 2D semantic guidance with 3D geometric features to generate accurate and spatially consistent 3D affordance masks. It consists of two key components: 1) Geometric Initialization, which projects 2D masks into 3D to produce coarse 3D mask candidates; and 2) a 3D Affordance Refinement module, which leverages local geometric context and self-attention to refine the masks.

Geometric Initialization. Given selected task-relevant frames \mathcal{V}^* and validated manipulable points $\mathcal{P}^{\text{valid}} \cup \mathcal{P}^*$, we first generate 2D binary masks for each frame $v_i \in \mathcal{V}^*$, denote as $\{\mathcal{M}_v^{2D}\}$ where $\mathcal{M}_v^{2D} \subseteq \{0, 1\}^{H \times W}$ using a pre-trained segmentation model. These masks are then projected into the 3D point cloud $\mathcal{PC} = \{p_i \in \mathbb{R}^3\}_{i=1}^M$ using known camera intrinsics $K \in \mathbb{R}^{3 \times 3}$ and extrinsics $[R | t] \in \mathbb{R}^{3 \times 4}$,

resulting in an initial 3D mask $\mathcal{M}_{\text{init}} \in \mathbb{R}^{M \times 1}$. This provides a coarse binary indicator for each 3D point’s relevance.

To tackle the challenges of processing high-fidelity scene point clouds due to their scale and precision, we focus on points within N -neighbors of those points masked by $\mathcal{M}_{\text{init}}$ to obtain $\hat{\mathcal{M}}^{3D}$. This design aligns with our coarse-to-fine framework.

3D Affordance Refinement. To capture local geometric dependencies in 3D space, we adopt the Point Transformer Layer inspired by (Zhao et al. 2021). This layer performs self-attention among neighboring points, enhanced by relative positional encoding. Given input point coordinates $p \in \mathbb{R}^{N \times 3}$ and features $x \in \mathbb{R}^{N \times C}$, the process is as follows:

First, the input features are projected into query, key, and value spaces via linear layers:

$$q_i = W_q x_i, k_j = W_k x_j, v_j = W_v x_j. \quad (6)$$

For each point p_i , we identify its k nearest neighbors $\{p_j\}_{j=1}^k$ and compute the relative positions $\delta_{i,j} = p_j - p_i$. These offsets are encoded by a multi-layer perceptron (MLP) to yield position embeddings:

$$\gamma_{i,j} = \text{MLP}_{\text{pos}}(p_j - p_i). \quad (7)$$

Then, we compute attention weights using a combination of feature difference and position encoding. Specifically, for each point pair (i, j) , the unnormalized attention weight is computed as:

$$w_{i,j} = \text{MLP}_w(k_j - q_i + \gamma_{i,j}), \quad (8)$$

where MLP_w is a shared-weight network applied across local neighbors. These weights are normalized via a softmax function:

$$\alpha_{i,j} = \frac{\exp(w_{i,j})}{\sum_{l=1}^k \exp(w_{i,l})}. \quad (9)$$

The output feature at each point is computed by aggregating the value features from its neighbors, weighted by attention and enriched with positional encoding:

$$x'_i = \sum_{j=1}^k \alpha_{i,j} \cdot (v_j + \gamma_{i,j}). \quad (10)$$

Leveraging the Point Transformer Layer, we employ an encoder-decoder architecture for refining 3D affordance masks, as showed in Figure 2. The encoder aggregates local geometric features and downsamples the point cloud to capture global contextual information. The decoder upsamples features via feature propagation layers with distance-weighted interpolation, while integrating skip connections to fuse encoder-derived features and retain fine-grained details, thereby ensuring precise segmentation boundaries.

Training Objective. To handle the challenges of mask prediction in high-resolution point clouds, we adopt a multi-objective loss function to supervise the training of the point transformer-based 3D refinement module. The total loss is defined as:

$$L_{\text{total}} = \alpha_1 L_{\text{BCE}} + \alpha_2 L_{\text{Dice}} + \alpha_3 L_{\text{Focal}} + \alpha_4 L_{\text{IoU}}, \quad (11)$$

Method	mAP	AP ₅₀	AP ₂₅	mIoU
OpenMask3D	0.2	0.2	0.4	0.2
LERF	0.0	0.0	0.0	0.0
OpenIns3D	0.0	0.0	0.0	0.1
Fun3DU	7.6	16.9	33.3	15.2
Ours	23.2	26.9	28.6	19.7

Table 1: Comparison of our method and baselines on the SceneFun3D dataset. All metrics are positively correlated with performance (higher is better).

Backbone	Model FLOPS	Inference Time (s/sample) \downarrow	Speed Up \uparrow
Fun3DU	48.05G	130.26	1.00 \times
Ours	29.50G	37.61	3.37\times

Table 2: Efficiency comparison between our method and Fun3DU on the SceneFun3D dataset. Inference time per task is reported using an NVIDIA A800 GPU.

where $\alpha_i \geq 0$ and $\sum_{i=1}^4 \alpha_i = 1$. Each loss term addresses a distinct aspect of the 3D segmentation quality: 1) L_{BCE} supervises per-point classification by minimizing pointwise prediction error. 2) L_{Dice} encourages overlap between predicted and ground-truth regions, promoting region-level consistency. 3) L_{Focal} emphasizes hard examples, mitigating class imbalance by down-weighting easy negatives. 4) L_{IoU} directly optimizes the intersection-over-union between predicted and ground-truth masks, enhancing spatial alignment. Balancing point-level precision and global coherence via 2D-3D priors, this loss yields robust affordance masks in dense scenes.

Experiment

Experimental Settings

Dataset. Evaluation conducted on SceneFun3D (Delitzas et al. 2024), the only existing dataset that provides annotations for 3D functional affordance segmentation in indoor environments. The dataset includes high-resolution 3D scans of real-world indoor scenes and is divided into two subsets: Split0, containing 30 scenes, and Split1, comprising 200 scenes. Each scene is accompanied by approximately 100 high-resolution RGB-D images, along with their corresponding intrinsic and extrinsic camera parameters. In addition, each scene is annotated with about 15 natural language task descriptions and their corresponding ground-truth 3D affordance masks. We report results on both splits to evaluate performance on small- and large-scale settings.

Evaluation Metrics. We evaluated our model using the standard metrics provided by SceneFun3D (Delitzas et al. 2024), which measure precision in 3D affordance segmentation. Specifically, we reported: 1) Average Precision (AP) at IoU thresholds of 0.25 and 0.5. 2) Mean Average Precision (mAP), computed by averaging AP across IoU thresholds

Affordance Weight	Double Check	mAP	AP ₅₀	AP ₂₅	mIoU
		3.9	6.00	14.9	4.4
✓		9.3	14.2	23.3	11.3
	✓	13.1	18.1	27.9	8.5
✓	✓	23.2	26.9	28.6	19.7

Table 3: Ablation of the Task-Aware 2D Affordance Detection module on SceneFun3D Split0.

geometry	color	text	mAP	AP ₅₀	AP ₂₅	mIoU
			7.4	18.7	29.9	16.7
✓	✓		19.3	19.3	24.3	15.2
✓		✓	9.2	10.0	10.1	4.8
✓	✓	✓	5.0	5.8	6.1	2.8
✓			23.2	26.9	28.6	19.7

Table 4: Ablation of the 3D Affordance Refinement Module with different input modalities: geometry, color, and text.

from 0.5 to 0.95 in 0.05 increments; 3) Mean Intersection over Union (mIoU), which measures per-point segmentation quality across the entire point cloud.

Baselines. We compared our approach against recent open-vocabulary 3D segmentation methods used in the SceneFun3D benchmark. Specifically, we included OpenMask3D (Takmaz et al. 2023), LERF (Kerr et al. 2023), OpenIns3D (Huang et al. 2024), as well as the more recent Fun3DU (Corsetti et al. 2025), which directly targets 3D scene affordance segmentation.

Implementation Details. We used Qwen (Bai et al. 2023) for both affordance concept extraction and the double-check mechanism. For coarse 2D mask generation, we combined MolMo (Deitke et al. 2025) with the Segment Anything Model (SAM) (Kirillov et al. 2023), leveraging their complementary strengths in vision-language alignment and generic segmentation. The 3D affordance refinement module is implemented using a Point Transformer (Zhao et al. 2021), trained with the multi-objective loss described in Eqn. (11). We set the loss weights to: $\alpha_1 = 0.3$, $\alpha_2 = 0.3$, $\alpha_3 = 0.2$, and $\alpha_4 = 0.2$. For geometric feature extraction, the number of N-nearest neighbors is set to 8,192.

Quantitative Results

Table 1 presents a comparison of our method with baselines on the SceneFun3D dataset. The results reveal that LERF and OpenIns3D achieve (AP₂₅) scores close to zero, while OpenMask3D attains a modest (AP₂₅) of 0.4. These limitations primarily arise from undersegmentation issues, such as misclassifying entire cabinets as handles, due to insufficient fine-grained semantic reasoning. These methods struggle with high-resolution point clouds, containing up to 13 million points against only a few hundred annotated ground-truth affordance masks, severely limiting their learning and resulting in low AP scores. Fun3DU achieves a notable AP₂₅

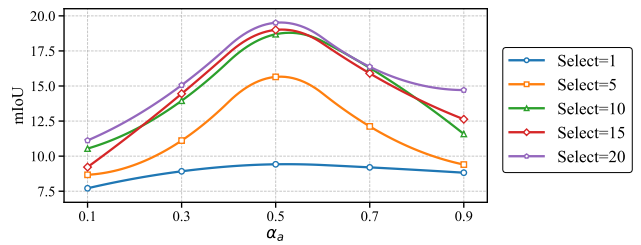


Figure 4: Effect of the number of selected images K on segmentation performance, evaluated using mIoU. Each experiment is conducted with a different Affordance Weight α_a .

score of 33.3, likely due to its ability to generate masks that satisfy the lenient mIoU threshold despite their imprecision. However, its performance is limited by directly translating 2D results into 3D without leveraging geometric structures, leading to a lower AP₅₀ of 16.9 and an mIoU of 15.2.

In contrast, our method demonstrates a significant advancement in precision, achieving the AP₅₀ of 26.9 and mIoU of 19.7, surpassing all baselines. This improvement is driven by our synergistic pipeline: 1) CLIP-driven, task-aware frame selection effectively filters irrelevant frames, thereby reducing noise and mitigating undersegmentation issues observed in LERF and OpenMask3D, leading to enhanced (AP) scores. 2) A coarse-to-fine integration of 2D semantic guidance with 3D geometric reasoning provides richer contextual understanding, resulting in superior (AP) and (mIoU) scores compared to Fun3DU.

Moreover, Table 2 quantifies the efficiency of our method compared to Fun3DU, achieving a 3.37 \times speedup with 40% fewer FLOPs. This efficiency stems from CLIP-driven task-aware frame selection, which eliminates irrelevant frames and redundant computations, enhancing inference throughput while maintaining performance. This efficient approach is more suitable for embodied downstream applications.

Ablation Study

On Task-Aware 2D Affordance Detection. We evaluated the Affordance-Weighted Frame Selection and Double-Check Point Validation components through ablation experiments as shown in Table 3. The model performs poorly without either component, while enabling each module individually yields moderate improvements. Specifically, the frame selection module raises mIoU to 11.3, while the Double-Check mechanism increases mAP to 13.1. Notably, activating both modules simultaneously yields the most significant gains across all metrics. These results underscore the complementary roles of the two components: the frame selection module enhances spatial priors by focusing on task-relevant views, while the Double-Check mechanism improves point-level accuracy through geometric validation. Together, they improve segmentation precision and robustness.

Figure 4 illustrates the impact of the number of selected K -frames. When only one highest-confidence frame is selected, performance is limited due to insufficient context. Although $K=20$ has the highest mIoU, $K=10$ achieves com-

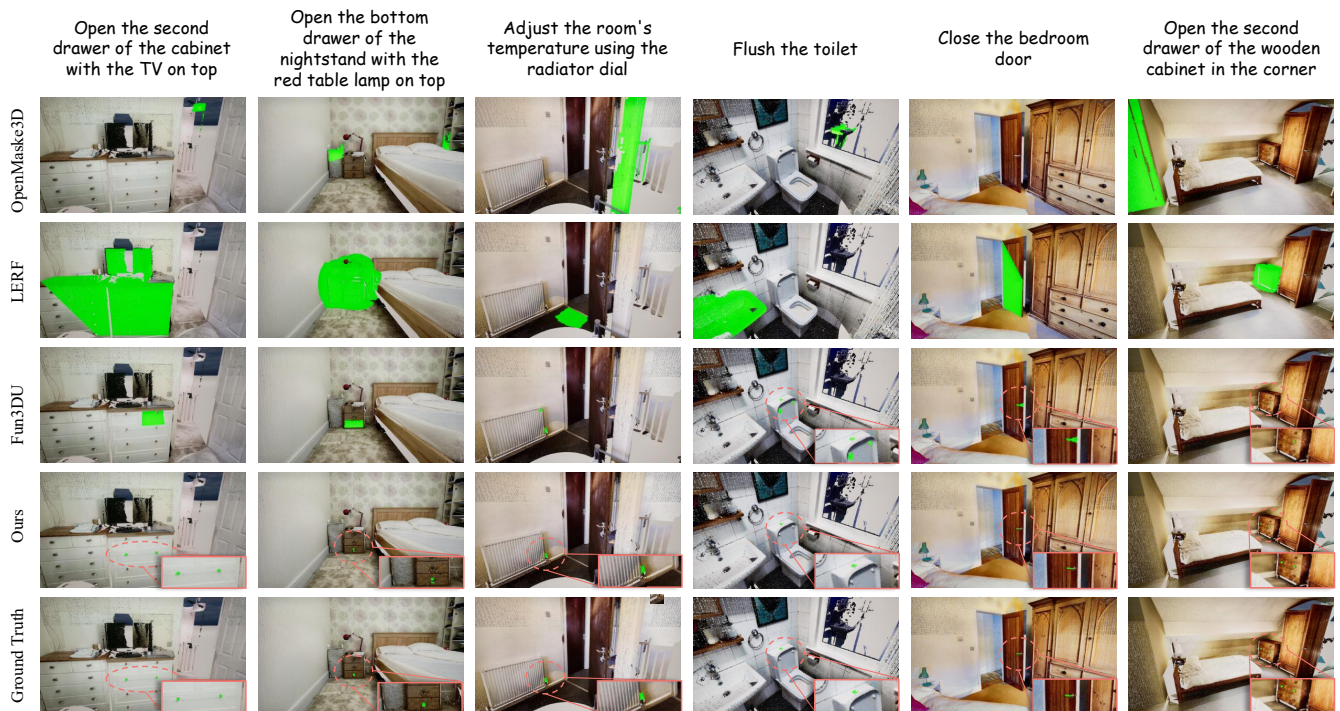


Figure 5: Qualitative comparison on SceneFun3D Split0. Point clouds are cropped around functional objects for improved visibility.

parable performance while effectively capturing critical visual cues. Therefore, we select $K=10$ as it offers a favorable trade-off between accuracy and computational efficiency.

On 3D Affordance Refinement. We conducted ablation experiments reported in Table 4. The results confirm that incorporating geometric structure is critical, yielding the highest scores in mAP, AP₅₀, and competitive performance in AP₂₅. We further explore the effect of adding color and text features to the 3D module. Surprisingly, performance degrades across all metrics when these modalities are included. We hypothesize that this decline results from overfitting or confusion introduced by low-quality or sparse semantic signals, especially under the limited data regime of SceneFun3D. These additional modalities may act as noise, hindering convergence and diluting the contribution of reliable geometric cues. This finding highlights the importance of modality alignment: adding more input channels does not guarantee better performance, especially when data is limited. Instead, carefully balancing semantic and geometric information is essential for effective affordance segmentation.

Qualitative Results

As shown in Figure 5, OpenMask3D often fails to interpret large-scale point clouds and struggles with basic scene understanding. For example, in the sixth column, it misclassifies upright wall features as part of a cabinet. In many cases, it is unable to localize the target object at all, likely due to its limited semantic grounding and computational inefficiency on dense inputs. LERF focuses on high-level object cate-

gories but lacks the ability to isolate task-relevant functional components. In the first and second columns, where the task involves “opening a cabinet drawer”, LERF highlights the cabinet itself while overlooking the drawer handle, illustrating its tendency to prioritize object names over actionable parts. Fun3DU performs better in identifying functional regions but suffers from spurious mask generation. In the first and second columns, it produces large, imprecise masks due to over-reliance on noisy 2D point predictions from VLMs. Additionally, when nearby textures are visually similar to the affordance region, as in the third and fourth columns, Fun3DU frequently oversegments irrelevant areas. It also struggles with tasks involving multiple disconnected components (e.g., dual handles), as seen in the sixth column.

In contrast, our method produces significantly more accurate and robust affordance segmentation results compared to existing baselines, particularly in complex and cluttered scenes. These improvements stem from the combination of 2D-guided localization and 3D geometry-aware refinement.

Conclusion

We present a Task-Aware 3D Scene-level Affordance segmentation framework (TASA), that jointly leverages 2D semantic cues and 3D geometric reasoning in a coarse-to-fine manner, marking the first geometry-optimized method for holistic affordance understanding in complex 3D scenes. The framework demonstrates strong performance to high-fidelity point clouds and consistently outperforms baselines in cluttered, real-world environments.

Acknowledgments

This research was funded by the Zhongguancun Academy (Grant No. 20240306), the National Natural Science Foundation of China (Grant No. 62376140, No. U23A20315 and No. 62376266), the Special Fund for Taishan Scholar Project of Shandong Province.

References

- Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, 422–440. Springer.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.
- Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, 202–221. Springer.
- Chu, H.; Deng, X.; Lv, Q.; Chen, X.; Li, Y.; Hao, J.; and Nie, L. 2025. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *arXiv preprint arXiv:2502.20041*.
- Corsetti, J.; Giuliani, F.; Fasoli, A.; Boscaini, D.; and Poesi, F. 2025. Functionality understanding and segmentation in 3D scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24550–24559.
- Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 91–104.
- Delitzas, A.; Takmaz, A.; Tombari, F.; Sumner, R.; Pollefeys, M.; and Engelmann, F. 2024. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14531–14542.
- Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; and Nießner, M. 2020. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9031–9040.
- Gibson, J. J. 2014. *The ecological approach to visual perception: classic edition*. Psychology press.
- Huang, Z.; Wu, X.; Chen, X.; Zhao, H.; Zhu, L.; and Lasenby, J. 2024. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *European Conference on Computer Vision*, 169–185. Springer.
- Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. Lurf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19729–19739.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.
- Li, G.; Sun, D.; Sevilla-Lara, L.; and Jampani, V. 2024. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3086–3096.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; and Chua, T.-S. 2018a. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 15–24.
- Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018b. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, 843–851.
- Lu, D.; Kong, L.; Huang, T.; and Lee, G. H. 2025. Geal: Generalizable 3d affordance learning with cross-modal consistency. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1680–1690.
- Luo, J.; Fu, J.; Kong, X.; Gao, C.; Ren, H.; Shen, H.; Xia, H.; and Liu, S. 2022. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16454–16463.
- Nguyen, T.; Vu, M. N.; Vuong, A.; Nguyen, D.; Vo, T.; Le, N.; and Nguyen, A. 2023. Open-vocabulary affordance detection in 3d point clouds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5692–5698. IEEE.
- Qian, S.; Chen, W.; Bai, M.; Zhou, X.; Tu, Z.; and Li, L. E. 2024. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7587–7597.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shao, Y.; Zhai, W.; Yang, Y.; Luo, H.; Cao, Y.; and Zha, Z.-J. 2025. Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17326–17336.

Sun, J.; Qing, C.; Tan, J.; and Xu, X. 2023. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2393–2401.

Takmaz, A.; Fedele, E.; Sumner, R. W.; Pollefeys, M.; Tombari, F.; and Engelmann, F. 2023. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*.

Van Vo, T.; Vu, M. N.; Huang, B.; Nguyen, T.; Le, N.; Vo, T.; and Nguyen, A. 2024. Open-vocabulary affordance detection using knowledge distillation and text-point correlation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 13968–13975. IEEE.

Wu, C.; Ma, Y.; Chen, Q.; Wang, H.; Luo, G.; Ji, J.; and Sun, X. 2024. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5940–5948.

Xu, L.; Gao, Y.; Song, W.; and Hao, A. 2024. Weakly supervised multimodal affordance grounding for egocentric images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6324–6332.

Xu, P.; and Mu, Y. 2025. Weakly-supervised affordance grounding guided by part-level semantic priors. *arXiv preprint arXiv:2505.24103*.

Yang, Y.; Zhai, W.; Luo, H.; Cao, Y.; Luo, J.; and Zha, Z.-J. 2023. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10905–10915.

Yang, Y.; Zhai, W.; Wang, C.; Yu, C.; Cao, Y.; and Zha, Z.-J. 2024. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems*, 37: 54529–54557.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.